Chemometrics

on its way towards

Personalized Health Care

Jasper Engel

CHEMOMETRICS ON ITS WAY TOWARDS PERSONALIZED HEALTH CARE

ISBN: 978-94-91602-46-7
Printing: Printservice Ede

CHEMOMETRICS ON ITS WAY TOWARDS PERSONALIZED HEALTH CARE

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus,
volgens besluit van het college van decanen
in het openbaar te verdedigen op maandag 21 maart 2016
om 14.30 uur precies

door

Jasper Engel

geboren op 30 september 1984
te Gouda

Promotoren:

Prof. dr. L.M.C. Buydens

Prof. dr. R.A. Wevers

Copromotoren:

Dr. L. Blanchet (*Universiteit Maastricht*)

Dr. U.F.H. Engelke

Manuscriptcommissie:

Prof. dr. A.J. van Gool (*voorzitter*)

Prof. dr. E.R. van den Heuvel (*Technische Universiteit Eindhoven*)

Dr. R. Wehrens (*Wageningen UR*)

Voor mijn ouders

Contents

1

GENERAL INTRODUCTION

Medical treatments can be quite successful for some individuals, while not for others. The rapid development of high-dimensional analytical laboratory technologies is expected to provide novel biomarkers or biomarker profiles for specific diseases. These techniques allow for examination of a patient at molecular level in unprecedented detail. This way, more personalized diagnosis and treatment plans may be applied. This is referred to as precision medicine. Data analysis is a crucial step in such personalized health care since the acquired data is often extremely large and complex. Unfortunately, many traditional methods have been designed for analysis of much smaller data sets and are not applicable. Therefore, development of novel data analysis approaches is essential to effectively use all the new information that is being gathered for the benefit of the patient.

This thesis seeks to address several issues in data analysis of high-dimensional technologies for personalized health care.

1.1 High-dimensional analysis in medicine

High-dimensional studies hold great promise in many fields such as the physical, chemical and health sciences [1, 2]. Advances in techniques such as DNA microarrays, infrared (IR) spectroscopy, liquid chromatography – mass spectrometry (LC-MS), and nuclear magnetic resonance (NMR) spectroscopy have enabled the rapid acquisition of a huge amount of molecular data in a sample in a single experiment [3, 4]. Subsequent application of powerful data analysis techniques allows for "data driven" research where measurements are performed to generate hypotheses rather than to confirm them [3].

High-dimensional techniques are becoming increasingly popular in medical applications [4-9]. Due to the advent of the so-called omics sciences it is now possible to study the molecular complexity of biological systems in unprecedented detail and at different levels ranging from the complete genome to profiles of metabolites [10]. The omics sciences have greatly improved our understanding of health and disease. For example, by means of data driven omics research many molecular profiles have been observed that correlate with specific biological processes and outcomes such as disease, thereby offering new information for processes whose mechanisms are yet unclear. Additionally, such signatures (biomarkers) can potentially be used as objective markers for e.g. diagnosis of a specific disease or selection and monitoring of therapy [11-14]. It is expected that the omics revolution will lead to significant advances in personalized health care where treatment is completely catered towards the unique molecular characteristics of an individual patient [4, 12, 13]. This is in great contrast to most current treatments that can be very successful for some individuals while not for others since they have been designed for the average patient with that specific disease. Moreover, the omics sciences have the potential to transform medicine from conventional reactive approaches (i.e. symptom-oriented diagnosis and treatment of diseases) towards disease prevention and early diagnostics [12].

1.1.1 What are the omics sciences?

Omics techniques are relevant for many clinical disciplines. They offer complementary views on human health and disease. The five major ones within the medical sciences are genomics, transcriptomics, proteomics, metabolomics, and glycomics [10]. These techniques are aimed at the universal detection of genes, mRNA, proteins, metabolites, and glycosylation profiles, respectively. The information in the genes is called the genotype. It is ultimately responsible for the final composition of a body fluid or tissue sample referred to as the biochemical phenotype. The genome offers much information regarding an individual's susceptibility to genetic diseases, and prediction of treatment response for a given patient. It may also help to elucidate molecular mechanisms of disease, identify

Box 1.1 Glossary

Biomarker
A biological marker (biomarker) is "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to therapeutic intervention" [14].

High dimensional studies
Experiments that are characterized by the measurement of a few dozen to many thousands of variables (dimensions) per sample. The resulting data from such experiments is often referred to a high-dimensional data. Due to the "curse of dimensionality" phenomenon many traditional data analysis methods do not scale well to high-dimensional data.

Omics
The collective name for technologies used to comprehensively study the molecules that make up a biological sample such as a cell, tissue, organ, body fluid, or organism. The techniques are primarily aimed at the universal detection of genes (genomics), RNA (transcriptomics), proteins (proteomics), metabolites (metabolomics), and glycosylation profiles (glycomics) in an untargeted and unbiased manner. Such analysis is sometimes referred to as high-dimensional biology.

Personalized health care
In this thesis, personalized health care is synonymous for precision medicine, which refers the use of molecular information (biomarkers) to target disease prevention and treatment to the needs of the individual patient. Outside of this thesis, personalized health care is sometimes viewed as a broader area that also includes concepts such as participatory health care, patient involvement, self-monitoring, etc.

Targeted analysis
A study that focuses on the measurement of a particular and limited set of compounds such as a particular set of metabolites.

Untargeted analysis
The simultaneous unbiased measurement of as many compounds (e.g. metabolites) as possible from a set of samples.

potential therapeutic targets, design drugs, etc [5]. However, a specific gene sequence is not directly related to a final disease outcome. This is because the genes are first transcribed and translated into proteins many of which can act as enzymes in numerous metabolic reactions. In addition to the genetic makeup of an individual, several environmental factors also influence these processes. This means that it is possible to observe patients with a similar genotype, but unique proteome (the complete set of proteins), metabolome (the complete set of metabolites), glycome composition, and also different disease outcome [7, 15]. Often, genomic information alone is not adequate to predict disease onset. This motivates the need to augment the information offered by genomics with the other omics

technologies to obtain information at the functional level of cells and tissues (i.e. the proteome, metabolome, and glycome). In concert these will influence the patient's clinical phenotype [5].

1.1.2 Metabolomics

The metabolome is the final downstream product of cellular activity and is thus an important driver of the clinical phenotype [16, 17]. It differs greatly between individuals and populations [18, 19]. The dynamics of metabolism operates on a timescale of seconds and depends on the interplay between genes and environmental factors such as diet, lifestyle and gut microbial composition [18]. These are the same factors that cause disease. Therefore, metabolomics is expected to become an important aspect of personalized health care as a sensitive and powerful method to study health and disease [4, 7, 15, 20, 21]. An additional advantage of metabolomics is that many biomarkers can be measured in a relatively noninvasive way by analyzing blood or urine samples [22]. Such samples are easily obtainable in the clinic. Below, three clinical examples of metabolomics studies are presented, which are used throughout this thesis. Note that these are all examples of so-called untargeted metabolomics studies (see box 1.1). In this thesis novel data analysis strategies are developed for such untargeted metabolomics studies.

Application 1: A nutritional intervention metabolomics study

Much effort is spent to understand how genetic and environmental factors influence the metabolome exactly [15]. This is required for a better understanding of health and disease, but also because these factors can act as confounders in clinical studies.

Nutrition can be beneficial but also detrimental for health. However, it is still unclear how an individual's diet relates to a specific health outcome exactly. In recent years, investigation of effects of nutrition on health and disease by metabolomics has gained significant interest [15]. The goal is to understand what happens to the metabolome with changes in diet to elucidate the underlying mechanisms and identify potential biomarkers. For example, a possible therapy of metabolic syndrome is to influence the metabolism by changing food intake patterns. In this respect it is of interest to understand how specific diets influence the metabolism and why some patients respond to a diet (i.e. lose a significant amount of weight) and others do not.

In chapter 4 of this thesis the metabolic response and loss of body weight due to four isocaloric diets was followed through time in individuals suffering from metabolic syndrome [23]. Urine ^{1}H NMR metabolomics was used to assess the metabolome at the

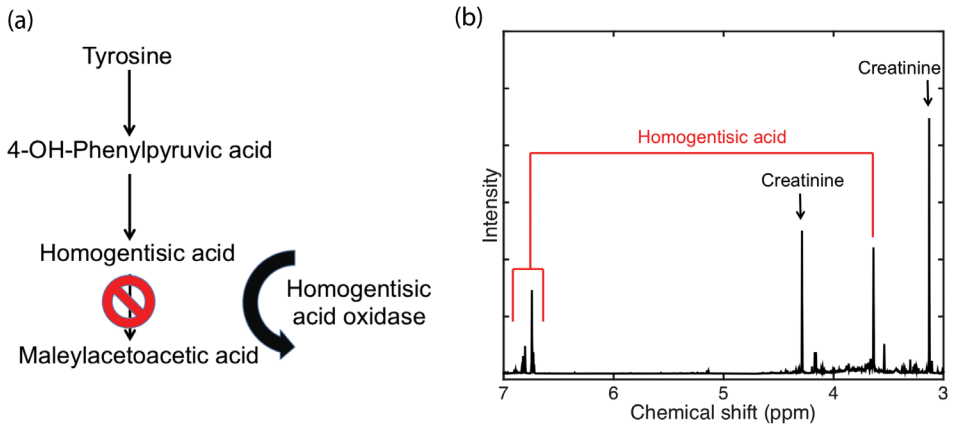

Figure 1.1 The metabolic pathway of tyrosine, showing the enzymatic defect present in alkaptonuria patients. **(a)** Tyrosine is metabolized to produce maleylacetoacetic acid in three steps. A defect of the enzyme homogentisic acid oxidase causes the build-up of homogentisic acid, and metabolites of homogentistic acid that do not occur in healthy individuals. These metabolites cause damage to heart valves, cartilage, and precipitate in organs e.g. as kidney stones. **(b)** High concentrations of homogentisic acid can be observed in urine ^{1}H-NMR spectra from affected patients.

different time points. Significant metabolic differences between responders and non-responders and between the four diets were identified by means of regularized multivariate analysis of variance (rMANOVA; a novel data analysis method introduced in this chapter).

Application 2: Diagnosis of inborn errors of metabolism

Metabolomics is becoming an increasingly popular tool for disease diagnosis. Typically, the value of a specific biomarker is measured in the patient and it is investigated if this value differs significantly from reference values. For example, metabolomics has been used this way to diagnose cardiovascular disease, diabetes, cancer, neurological diseases, and inborn errors of metabolism (IEM) [20, 24-26].

IEM comprise a substantial group of rare genetic diseases. An IEM is often a defect in a single gene that leads to impaired activity of an enzyme, co-factor, or other relevant protein [27]. This often results in a block in a particular metabolic pathway leading to the accumulation of a metabolite before the block (e.g. the substrate of an enzyme) and/or a lack of one or more metabolites downstream of the block (the product of the enzymatic reaction). Alternatively, it can be that the block causes alternative pathways to be used resulting in accumulation of unexpected and possibly toxic metabolites. As an example, the diagnosis of Alkaptonuria by means of urine ^{1}H NMR metabolomics is shown in figure 1.1. The peaks at 3.64 and 6.76 ppm indicate an elevated concentration of homogentisic acid, a

biomarker for this disease, in the urine of the patient [25]. In chapters 5 and 6 of this thesis, novel data analysis methods are introduced that can be used to screen the metabolome of patients and automatically identify (patterns of) metabolite concentrations that are not expected to occur in healthy individuals. Subsequently, this information can be used for disease diagnosis. The methods are used to diagnose different IEM in a number of patients.

Application 3: Health monitoring by personalized omic profiling

Recently, metabolomics in combination with other omics technologies was used in a pioneering study to monitor a 54-year-old, initially healthy male, at 20 time points over a 14-months period [28]. The approach was coined integrated personal omic profiling (iPOP). During the period of the study the individual suffered from two episodes of viral infection. Temporal changes in the personal omics profile (based on measurements of blood components) showed strong association with health and disease status; for example, changes in metabolite expression levels appeared to be consistent with the emergence of the viral infections [4]. Additionally, genomics measurements, performed at the start of the study, revealed that the individual was predisposed to type-2 diabetes whose onset was observed immediately after one of the viral infections. Due to the early diagnosis offered by iPOP, the patient could effectively control and reverse the diabetes phenotype by proactive interventions such as a diet change and physical exercise [12].

The iPOP study demonstrated the potential of omics offering a very detailed assessment of someone's health status and closely monitoring disease onset. Temporal iPOP studies may become feasible in the near future with decreasing costs of omics experiments. A single assessment, however, is already doable and could be used to screen individual patients and potentially identify the root cause of disease as we show in chapters 5 and 6 of this thesis. The data analysis methodology introduced in these chapters can also be used to automatically inspect iPOP data from a temporal assessment. This is discussed in chapter 7.

1.2 Data analysis challenges in metabolomics

Due to the complex nature of the metabolome and the high-dimensional analytical techniques used to measure it, datasets are extremely large and complex. This is illustrated in figure 1.2 where a urine ^{1}H NMR spectrum of a healthy volunteer is shown. The complexity of this spectrum is, for example, clearly reflected by the large number of peaks that can be observed; the greatly varying peak intensities, which are related to the different metabolites in the sample; and the presence of signal that is due to noise. Factors such as baseline drift; peak shifts; and dilution differences between samples add further complexity

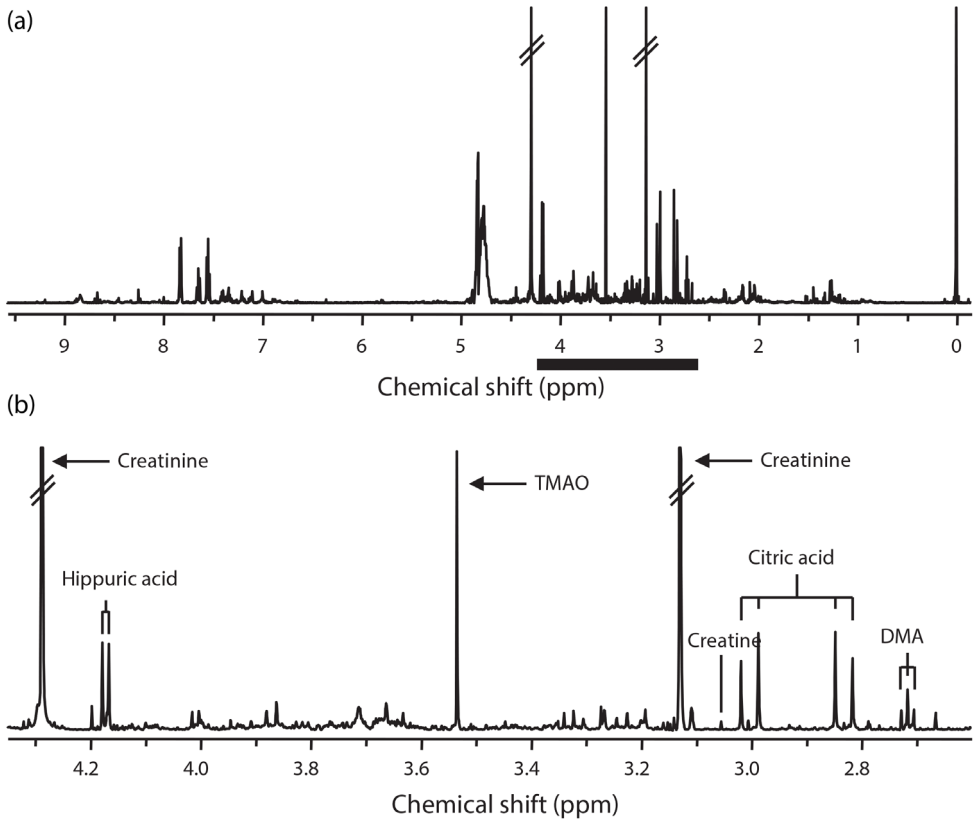

Figure 1.2 **(a)** ^{1}H-NMR spectrum of urine from a healthy volunteer and **(b)** an expansion of a part of this spectrum. DMA = dimethylamine; TMAO = trimethylamine N-oxide.

to the data. The complexity increases even further when other techniques such as LC-QTOF are used where more than 10000 peaks can be measured for each sample. Clearly, data analysis is indispensable for analysis of such large data [24, 29]. Ideally, a data analyst is involved in every step of a metabolomics experiment starting with its design (formulation of focused research questions, number of samples, sample collection, collection moments, study design, etc.), subsequent pre-processing of the measured data to remove unwanted artifacts, (multivariate) statistical modeling to extract the biological information, and statistical and biological validation of the results.

Analysis of metabolomics data is not straightforward. Currently the development of novel metabolomics technologies greatly outpaces our ability to properly analyze the acquired data. This is in a large part due to the large number of variables that are measured in a single experiment. Here we present three examples of data analysis challenges that are relevant for

this thesis. Note that many of these challenges are general challenges that also apply to data from other high-dimensional technologies such as proteomics or glycomics.

1. **Data pre-processing** is a crucial step to remove unwanted variation from the data such that subsequent analysis can focus on the relevant biological information in the data (e.g. the biomarker). For example, peak shifts in NMR spectra and chromatograms are usually corrected for by alignment or warping methods. Often it is not directly clear which method will perform best (in some relevant sense) for a specific data set and several settings and methods have to be investigated. However, many methods are computationally expensive when applied to data with a large number of variables. Therefore, this is not a viable approach for metabolomics. Guidelines for efficient data pre-processing have to be developed, ideally as a standard operation procedure.

2. Analysis of metabolomics data heavily relies on **statistical modeling**, for example, to detect metabolic differences between a group of healthy controls and a group of patients in a biomarker discovery experiment. Often, however, the number of samples in a data set is much smaller than the number of variables and many more or less standard statistical methods become extremely unreliable. One way to circumvent these issues is to impose some structure to the model with the aim to reduce variance at the cost of some bias. This is known as regularization [2, 30-33]. Dimension reduction by principal component analysis (PCA) (or a similar method related to explained variance) is a popular regularization approach in metabolomics. However, dimension reduction by PCA may not work out well, e.g. when group separation is only visible in the last (discarded) principal components. Clearly, other regularization approaches for analysis of metabolomics data have to be used in some applications.

3. **Data analysis for personalized health care** requires statistical modeling beyond population-based approaches. Metabolomics studies used to compare groups of samples to each other (e.g. patients with a specific disease and controls). However, for a number of reasons this may be impractical in a clinical setting. For example, in the case of diagnosis of rare diseases typically too few samples from diseased patients are available for a useful comparison. The population-based approach is also not useful for health monitoring (see application 3), because (1) it ignores the potential longitudinal aspect of such studies, and (2) the health status of an individual should be assessed in an unbiased way instead of focusing on one specific disease. Additionally, by definition, population-based models are designed

to be the best for the average population, but not necessarily for specific individuals. As mentioned above, individuals are genetically diverse and may have different exposures and response characteristics to disease [34]. These individual responses are likely to overlap in large population models [4].

It is clear that the new techniques require that we compare an $n = 1$ patient sample to a group of controls. With this paradigm shift something can be said about the individual patient. Ideally, a data analysis method that can screen the metabolome of individual patients and automatically identify abnormal (patterns of) metabolite concentrations would be extremely useful.

The three aforementioned examples demonstrate that further development of data analysis approaches is crucial to unlock the clinical potential of metabolomics (and other high-dimensional technologies) for personalized health care. For this purpose it is extremely interesting to consider novel approaches that have been proposed in fields outside of metabolomics. With respect to data pre-processing (challenge 1) the current and emerging approaches described in literature should be carefully evaluated on benchmark data. Based on these findings more efficient pre-processing guidelines can hopefully be developed. Statistical modeling of high-dimensional data for disease diagnosis, biomarker discovery, etc. (challenge 2) may benefit from the novel regularization approaches that have been developed in fields such as statistics, statistical learning, and machine learning [30, 32]. These developments bring interesting new perspectives and deserve more attention from the metabolomics community than they have received so far. In many situations these methods offer more powerful ways of regularization compared to dimension reduction with PCA. Finally, statistical modeling for personalized health care (challenge 3) may benefit from ideas that have been developed in the field of industrial process monitoring. Here, the goal is to automatically determine whether or not a process is in control (Normal Operating Conditions); if not, a known or unknown rare event has occurred that may affect product quality. This is essentially the same goal as monitoring the health status of an individual for early diagnosis of (rare) diseases.

1.3 Scope and outline of this thesis

This thesis addresses several challenges in pre-processing and statistical modeling of data generated by high-dimensional technologies (see box 1.2). The focus is on several topics in personalized health care, namely biomarker discovery, disease diagnosis, and health monitoring.

Box 1.2 Aims of this thesis

Data pre-processing

- Evaluation of current approaches for selection of an optimal method for pre-processing of a high-dimensional data set as a first step towards development of stable and effective guidelines.

Statistical modeling

- Development of novel regularized models for personalized healthcare that are in principle generally applicable to high-dimensional data sets originating from e.g. metabolomics, glycomics, or proteomics experiments.
- Improve e.g. biomarker discovery by development of a multivariate method for analysis of data with an underlying experimental design.
- Development of a multivariate model for the comparison of an $n = 1$ patient sample to data of a group of controls, e.g. for diagnosis of rare diseases or health monitoring.

In **Chapter 2** current strategies and guidelines for selection of the optimal pre-processing method are critically evaluated. It is shown that breaking with current trends in pre-processing is necessary: all selection approaches have serious drawbacks and are unreliable. Throughout the chapter Fourier transform-infrared (FT-IR) data, originating from a food-authentication study, is used to demonstrate the behavior of the approaches. This type of data set is well understood and therefore serves as a perfect benchmark. Metabolomics data is often more complicated, but conclusions drawn in the chapter definitely generalize to metabolomics data.

In **Chapter 3** modern regularization approaches for statistical modeling of high dimensional data are reviewed. It is shown that the multivariate techniques in this thesis are all related to the Mahalanobis distance and therefore to each other. Because of this, regularization is discussed in the context of this distance. However, many of the approaches described in this chapter are general and can also be used for other regularization of other methods. Chapter 3 provides the basis for the studies presented in chapters 4 – 6.

In **chapter 4** analysis of metabolomics data with an underlying experimental design is explored. Such data is often obtained in biomarker discovery experiments. Multivariate analysis of variance (MANOVA) is a traditional method for analysis of such data. However, it cannot be applied to high-dimensional data. ANOVA simultaneous component analysis (ASCA) is a recently proposed alternative to MANOVA for analysis of such data. We show that ASCA implicitly assumes that the variables in the data are uncorrelated. This unrealistic assumption reduces the power of the method and hampers interpretation. An improved model is proposed that is essentially a weighted average of the ASCA and MANOVA

models. This is achieved by combining MANOVA with a modern regularized estimator of the covariance matrix. The new method is used to analyze data from an untargeted metabolomics nutritional intervention study.

In **chapters 5 and 6** novel statistical approaches for identification of the disease biomarker in individual patients are introduced. For this purpose, ideas from industrial process monitoring are adapted to the clinical setting. The resulting methods can be used for analysis of temporal data (health monitoring) or a single assessment (e.g. to diagnose rare diseases). The latter approach is used in these chapters to diagnose several IEM. In **chapter 5** one of the most commonly used models for industrial process monitoring is transformed to the metabolomics setting. This model uses dimension reduction via PCA for regularization. As a consequence, the technique is computationally inexpensive, but reliable identification of the disease biomarker is not guaranteed in all situations. Therefore, a novel method is developed in **chapter 6**. More specifically, model estimation is combined with variable selection instead of dimension reduction. A so-called ℓ_1-norm constraint is used for this purpose. Because of this, a fast algorithm can be used to solve the problem. Additionally, the solution can be graphically depicted and provides and intuitive way for a clinical practitioner to combine expert knowledge with the output of the model. Note that this constrained approach to variable selection has barely been used in metabolomics so far. It is shown that the resulting method can often identify the individual biomarker more reliably compared to the method developed in chapter 5 at the cost of computational complexity.

In **chapter 7** the main findings of the research presented in this thesis are summarized. Next, various perspectives for further research are presented. Of all interesting avenues for further research, this chapter mainly focuses on application of the proposed data analysis methods to longitudinal metabolomics studies such as biomarker discovery and health monitoring.

References

1. Fan, J., F. Han, and H. Liu, *Challenges of Big Data analysis.* National Science Review, 2014.
2. Fan, J. and J. Lv, *A Selective Overview of Variable Selection in High Dimensional Feature Space.* Statistica Sinica, 2010. **20**(1): p. 101-148.
3. Buydens, L., *Towards tsunami-resistant chemometrics.* The analytical scientist, 2013. **813**: p. 24-30.
4. Li-Pook-Than, J. and M. Snyder, *iPOP goes the world: integrated Personalized Omics Profiling and the road towards improved health care.* Chemistry & biology, 2013. **20**(5): p. 660-666.
5. Jayachandran, D., et al., *Revitalizing Personalized Medicine: Respecting Biomolecular Complexities Beyond Gene Expression.* CPT: pharmacometrics & systems pharmacology, 2014. **3**(4): p. e110.
6. Dunn, W.B., N.J.C. Bailey, and H.E. Johnson, *Measuring the metabolome: current analytical technologies.* Analyst, 2005. **130**(5): p. 606-625.
7. Ellis, D.I., et al., *Metabolic fingerprinting as a diagnostic tool.* 2007.
8. Lenz, E.M. and I.D. Wilson, *Analytical Strategies in Metabonomics.* Journal of Proteome Research, 2006. **6**(2): p. 443-458.
9. Lindon, J.C., J.K. Nicholson, and E. Holmes, *The handbook of metabonomics and metabolomics*. 2007, Amsterdam, The Netherlands: Elsevier.
10. Horgan, R.P. and L.C. Kenny, *'Omic'technologies: genomics, transcriptomics, proteomics and metabolomics.* The Obstetrician & Gynaecologist, 2011. **13**(3): p. 189-195.
11. Poste, G., *Bring on the biomarkers.* Nature, 2011. **469**(7329): p. 156-157.
12. Chen, R. and M. Snyder, *Promise of personalized omics to precision medicine.* Wiley Interdisciplinary Reviews: Systems Biology and Medicine, 2013. **5**(1): p. 73-82.
13. Chen, R., et al., *Personal omics profiling reveals dynamic molecular and medical phenotypes.* Cell, 2012. **148**(6): p. 1293-1307.
14. Colburn, W., et al., *Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. Biomarkers Definitions Working Group.* Clinical Pharmacol & Therapeutics, 2001. **69**: p. 89-95.
15. McNiven, E.M., J.B. German, and C.M. Slupsky, *Analytical metabolomics: nutritional opportunities for personalized health.* The Journal of nutritional biochemistry, 2011. **22**(11): p. 995-1002.
16. Mamas, M., et al., *The role of metabolites and metabolomics in clinically applicable biomarkers of disease.* Archives of toxicology, 2011. **85**(1): p. 5-17.
17. Fiehn, O., *Metabolomics–the link between genotypes and phenotypes.* Plant molecular biology, 2002. **48**(1-2): p. 155-171.
18. Holmes, E., et al., *Human metabolic phenotype diversity and its association with diet and blood pressure.* Nature, 2008. **453**(7193): p. 396 - U50.
19. Assfalg, M., et al., *Evidence of different metabolic phenotypes in humans.* Proceedings of the National Academy of Sciences, 2008. **105**(5): p. 1420-1424.
20. Nicholson, J.K., et al., *Metabolic phenotyping in clinical and surgical environments.* Nature, 2012. **491**(7424): p. 384-392.
21. Holmes, E., I.D. Wilson, and J.K. Nicholson, *Metabolic Phenotyping in Health and Disease.* Cell, 2008. **134**(5): p. 714-717.
22. Beckonert, O., et al., *Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts.* Nature protocols, 2007. **2**(11): p. 2692-2703.
23. Soenen, S., *Efficacy of macronutrients on targeting obesity and beyond*, 2010, Universiteit van Maastricht.
24. Madsen, R., T. Lundstedt, and J. Trygg, *Chemometrics in metabolomics—A review in human disease diagnosis.* Analytica Chimica Acta, 2010. **659**(1–2): p. 23-33.
25. Engelke, U.F.H., et al., *Handbook of ^{1}H-NMR spectroscopy in Inborn Errors of Metabolism: Body fluid NMR spectroscopy and Brain MR spectroscopy*. 3 ed. 2014: Heilbronn: SPS Verlagsgesellschaft.
26. Lindon, J.C., E. Holmes, and J.K. Nicholson, *Metabonomics in pharmaceutical R & D.* FEBS Journal, 2007. **274**(5): p. 1140-1151.
27. Lanpher, B., N. Brunetti-Pierri, and B. Lee, *Inborn errors of metabolism: the flux from Mendelian to complex diseases.* Nature Reviews Genetics, 2006. **7**(6): p. 449-459.
28. Chen, R., et al., *Personal Omics Profiling Reveals Dynamic Molecular and Medical Phenotypes.* Cell, 2012. **148**(6): p. 1293-1307.
29. Trygg, J., E. Holmes, and T. Lundstedt, *Chemometrics in Metabonomics.* Journal of Proteome Research, 2006. **6**(2): p. 469-479.
30. Hastie, T.T., R.; Friedman, J., *The Elements of Statistical Learning*. 2001, New York, NY, USA: Springer New York Inc.

31. Pourahmadi, M., *Covariance Estimation: The GLM and Regularization Perspectives.* Statistical Science, 2011. **26**(3): p. 369-387.
32. Pourahmadi, M., *High-Dimensional Covariance Estimation*. Wiley Series in Probability and Statistics, ed. D.J.C. Balding, N.A.C.; Fitzmaurice, G.M.; Goldstein, H.; Johnstone, I.M.; Molenberghs, G.; Scott, D.W. 2013, Hoboken, New Jersey, US: John Wiley & Sons.
33. Bickel, P., et al., *Regularization in statistics.* Test, 2006. **15**(2): p. 271-344.
34. Johnson, J.M., et al., *A practical approach to detect unique metabolic patterns for personalized medicine.* The Analyst, 2010. **135**(11): p. 2864-2870.

2

BREAKING WITH TRENDS IN PRE-PROCESSING?

Data pre-processing is an essential part of chemometric data analysis, which aims to remove unwanted variation (such as instrumental artefacts) and thereby focusing on the variation of interest. The choice of an optimal pre-processing method or combination of methods may strongly influence the analysis results, but is far from straightforward, since it depends on the characteristics of the dataset and on the goal of data analysis. This chapter is devoted to the selection procedure of appropriate pre-processing strategies. We show that breaking with current trends in pre-processing is essential, as all selection approaches have serious drawbacks and cannot be properly used.

This chapter has been adapted from: Engel, J.[#], Gerretzen, J.[#], Szymańska, E., Jansen, J. J., Downey, G., Blanchet, L., & Buydens, L. M. C. (2013). Breaking with trends in pre-processing?. *TrAC Trends in Analytical Chemistry*, *50*, 96-106.

[#] These authors contributed equally to this study.

2.1 Introduction

Due to recent developments in analytical chemistry, increasing amounts of complex analytical chemical data are being produced and employed in many different application areas, ranging from metabolomics to industrial process monitoring and quality control [1, 2]. Such comprehensive data requires that increasingly more time and effort is invested in data analysis and interpretation, to extract the information of interest. Pre-processing is an important first step of data analysis (see figure 2.1) in which raw data is transformed to "cleaned" data, from which unwanted variation such as instrumental and experimental artefacts have been removed, so that this cleaned data is better suited to the data analysis goals [3, 4]. If not performed in the right way, pre-processing can also introduce unwanted variation. Proper pre-processing is thus a critical step that directly influences the successful outcome of all following steps in the pipeline and therefore the success of the entire experiment.

Pre-processing methods are mainly used to adjust the variability of each measured variable and their relationships (e.g. by normalisation, scaling) to better comply with the data analysis goal. Moreover, they deal with challenging and detrimental data characteristics, such as missing values and "data artefacts", which may be highly specific to the analytical chemical techniques used, such as baseline shifts in spectroscopy and peak shifts in chromatographic or NMR data [5-7]. The emerging use of hyphenated techniques, combinations of techniques (data fusion) and combining data collected in different time periods/analysis batches (batch-to-batch effects) or different laboratories (inter-laboratory comparisons) make these data artefacts an increasingly challenging issue [8].

The choice of an optimal pre-processing method or combination of methods depends on many different properties of the data and on the goal of data analysis. This choice requires very careful thought: when dealing with large data sets, the effects of each numerical data pre-processing step are not transparent. However, conclusions drawn should be robust to the type of data pre-processing that was applied [9].

Many different pre-processing methods have been developed [10-13]. It is still largely impossible to robustly predict whether a certain pre-processing method brings the data analysis goal closer. There are no clear-cut guidelines when to use or avoid certain pre-processing methods.

This chapter critically evaluates current approaches to selection of pre-processing. First, several common data artefacts for a variety of different analytical techniques are described,

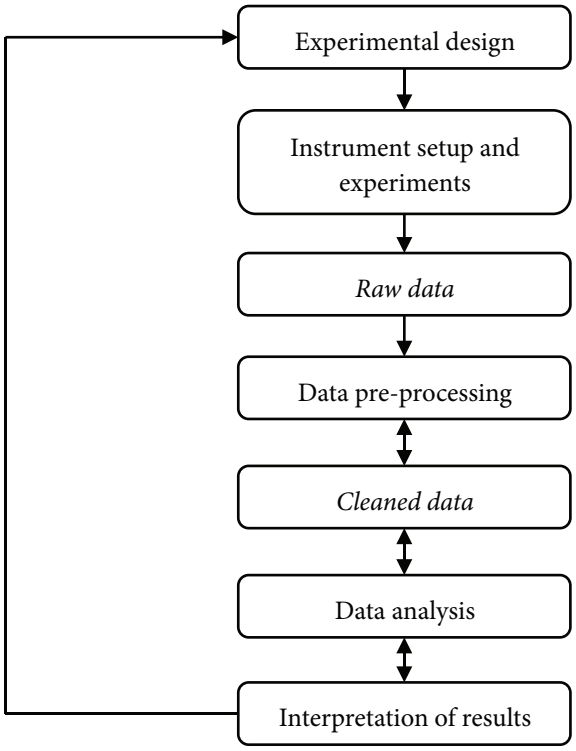

Figure 2.1 An overview of the pipeline for design, performance and analysis of experiments.

This chapter critically evaluates current approaches to selection of pre-processing. First, several common data artefacts for a variety of different analytical techniques are described, as well as the pre-processing methods available to reduce or remove these artefacts. Second, the approaches used to determine an optimal pre-processing method or strategy are discussed. A critical comparison between these three approaches is illustrated by the pre-processing of a mid-infrared (FTIR) data set with almost 5000 sensible combinations of pre-processing methods. Finally, we provide an outlook to the near future on where selection of pre-processing strategies should head next.

2.2 Data artefacts and pre-processing

Pre-processing data generally consists of multiple steps, each one correcting for a particular artefact. Several individual pre-processing "methods", each relevant to a specific artefact, will need to be applied consecutively in a pre-processing "strategy" to counter all artefacts

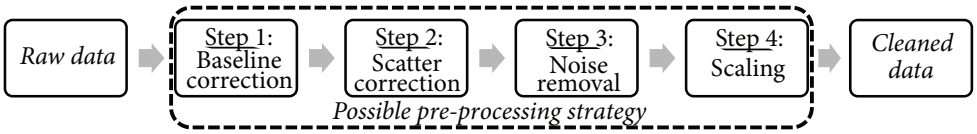

Figure 2.2 Schematic representation of a complete pre-processing strategy for IR data. Steps can be skipped or added and the order of steps can be changed, based on the data set at hand and the data analysis goal.

present in the data. Figure 2.2 shows a typical example of such a pre-processing strategy for the IR data used below for illustration. This section gives an overview of the data artefacts that are most commonly encountered in chemical data and frequently-used pre-processing methods that are available to correct for these artefacts.

The order in which the pre-processing methods are applied in the strategy may be specific to the data characteristics and the data analysis goal. Note that figure 2.2 presents only one out of many possible strategies: changing the order of the pre-processing steps may change the final results. Other steps often considered as pre-processing steps are outlier detection, feature selection, signal deconvolution and dealing with missing values. However, these steps are so entangled with the subsequent data analysis steps in figure 2.1 that they are not discussed further here.

It is our aim to provide a broad overview of data artefacts and analytical techniques, although not every artefact for each possible analytical technique will be discussed. However, the artefacts discussed in the following sections are mostly relevant to multiple analytical techniques. Also, some artefacts related to the sample composition instead of to the analytical technique are described. Reviews that describe artefacts and pre-processing methods for one particular analytical technique have been published elsewhere [4, 14-18]. Table 2.1 presents an overview of the most common artefacts for a set of different analytical techniques. Note that very different underlying physical phenomena—relevant to different analytical techniques—may cause very similar artefacts that may therefore be solved by the same pre-processing method.

Many dedicated software tools, such as MATLAB, R, The Unscrambler or SIMCA [19], can be used for data pre-processing. Some pre-processing methods are accessible via (commercially available) toolboxes or extensions of this software. Software supplied with analytical instruments often also contains some basic pre-processing functions. However, many researchers seem to use in-house written routines.

Table 2.1 Overview of data artefacts for the most common analytical data types. A cross (X) indicates that the artefact is present for an analytical data type.

Artefact	Spectroscopy (UV/IR[a]/Raman)	NMR	Chromato-graphy	Electro-phoresis	Mass-spectrometry	Fluorescence
Baseline	X	X	X	X	X	X
Misalignment		X	X	X	X	X
Scatter	X					
Noise	X	X	X	X	X	X

[a] IR includes both NIR and mid-IR data

2.2.1 Noise

Noise is common to almost any analytical technique and the underlying background differs per analytical technique. The principle of removing noise is commonly based on smoothing of a signal via, for example, the Savitzky-Golay algorithm or by using wavelets [20, 21]. The Savitzky-Golay algorithm is based on fitting polynomials to many small data windows (one polynomial per window). The wavelet approach relates to a Fourier-like transformation of the input signal where noise is removed by filtering the high frequency parts of the transformed signal – more details can be found in [21]. Other types of noise, like structured (*i.e.* non-white) technical noise, are not considered further here.

2.2.2 Baseline offset and slope

Baseline effects lead to signals having a vertical offset or a slope. Just as with noise, the underlying background for a baseline offset or slope differs per analytical technique. Figure 2.3 shows a baseline offset and a baseline slope for an artificial IR spectrum. Commonly, a baseline is corrected by estimation of the baseline and subsequent subtraction of the estimate from the measured spectrum.

Common methods for estimating a baseline are "detrending" [10] or estimation via "Asymmetric Least Squares Smoothing" (AsLS, [22]). In detrending, one fits a polynomial of a fixed degree to the spectrum and subsequently subtracts this polynomial from the spectrum. In AsLS, a smoother is combined with asymmetric weighting of deviations from the smoothed signal (*i.e.* peaks) to estimate the baseline.

Baseline effects can also be corrected by taking a derivative of the input signal. A first order derivative will eliminate a constant baseline (offset) and a second order derivative will also eliminate a baseline slope. Since derivatization can reduce the signal-to-noise ratio, the

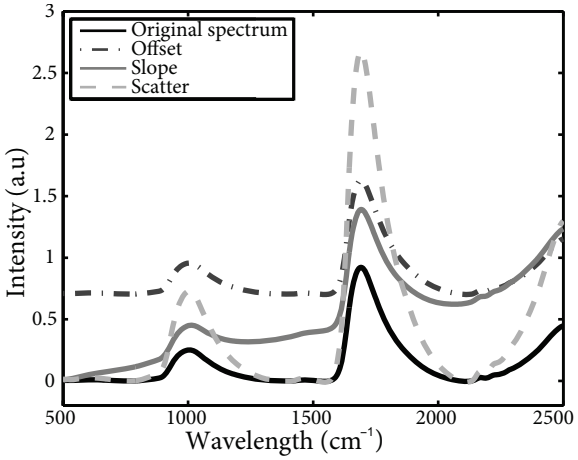

Figure 2.3 Examples of different artefacts in spectroscopic data. The solid line represents an original (artificial) IR spectrum. The others are modified versions of this original spectrum (dashed line: offset; dotted line: slope; dash-dotted line: scatter effect).

Savitzky-Golay algorithm described in section 2.2.1 is frequently used for this purpose, by calculating the derivatives of the fitted polynomials.

2.2.3 Light scatter

Scatter effects (see figure 2.3) are common to all analytical techniques that involve the use of light, such as IR or UV spectroscopy. Light scattering occurs because the size of the particles in the sample has at least one dimension that is roughly the same magnitude as the spectroscopic wavelengths. In most cases, scatter effects are corrected by comparing signal intensities to a reference signal. Here, the assumption is made that most systematic differences between reference and the spectrum to correct are due to scattering. The size of the scatter effect is described by a constant multiplied by the intensity of the spectrum. All spectra are corrected for scattering by dividing them by their estimated scatter constant.

Many pre-processing methods can correct for light scatter effects. For example, Standard Normal Variate (SNV, [10]) and RNV (Robust Normal Variate, [23]) are methods that subtract the spectrum mean (or median for RNV) from each spectral variable and subsequently divide that value by the (robust) standard deviation of the spectrum (*i.e.* the estimated scatter constant). Alternatively, Multiplicative Signal Correction (MSC, [24]) tries to estimate the coefficient describing the scattering by regressing the spectrum to correct on to a reference. The close relationship between MSC and SNV is elaborated on elsewhere [25].

2.2.4 Temporal and spectral misalignment

Environmental fluctuations in the sample, laboratory and the analytical technology can influence the expression of the same measured sample in the spectral dimension (*e.g.*, NMR chemical shift, due to pH changes) or in the temporal dimension (*e.g.*, chromatographic time shifts).

These differences are commonly solved by applying "alignment" methods (also known as "warping" methods) to the data [26, 27]. Most alignment methods stretch or compress each signal to match a certain reference signal in the best possible way. Some methods, like Parametric Time Warping (PTW, [22]), aim for a global alignment model, such that the stretching (or compression) can be described by one so-called warping function (*i.e.* the full signal is warped at once). Such alignment is deemed most suitable for chromatographic data.

Other methods, like Correlation Optimised Warping (COW, [12]) and *i*coshift [28], apply local alignment. These methods split the signal into different segments and optimal alignment should be reached by stretching or compressing the individual segments to match the segments of the reference best. Local alignment is particularly relevant for the pre-processing of NMR spectra where each peak may shift in both (spectral) directions. Additionally, the amount of shifting can vary between peaks.

*i*coshift was originally proposed for NMR data, but has been successfully applied for chromatographic misalignment as well [29], clearly showing that the same pre-processing methods may benefit analytical artefacts with very different physical backgrounds.

Alignment methods suffer from two limitations. The first one is purely technical: the algorithms are relatively slow, making parameter optimisation a difficult task. Moreover, these time-intensive calculations limit their applicability to large-scale datasets, such as those encountered in metabolomics. Second, other analytical artefacts present in the data, such as baseline effects, scatter effects or high noise levels considerably hamper the performance of most alignment methods. However, several methods that tackle these other artefacts may rely on properly aligned data, again indicating that the choice of an optimal pre-processing strategy is not straightforward (figure 2.2).

An alternative approach for correcting (small) misalignments is binning, also known as bucketing, which is routinely applied to NMR data [17]. In binning, the signal is split into many segments (called bins). The integral of the signal, or, alternatively, the maximum intensity in each bin is used as a replacement for the original signal, thereby reducing the

effect of small misalignment [17]. An additional advantage of binning is reduction of data size, which may facilitate data analysis by reducing computation time. Consequently, the resolution of the spectra is reduced. Important in binning is placing the bin boundaries to avoid splitting a peak among multiple bins, so sophisticated binning methods that take into account the position of peaks in an NMR spectrum have been devised. An example can be found in [30].

Note that misalignment in data does not necessarily have to be corrected prior to data analysis. For example, PARAFAC2 can implicitly deal with misalignment in one data mode (*e.g.* in the chromatographic mode of LC-MS data) [31].

2.2.5 Normalization, scaling and element-wise transformations

The previously described artefacts are all related to the analytical techniques used to produce the data. However, artefacts can also be directly related to the sample under study.

A good example can be found in NMR spectra of urine [32], in which the area of an NMR peak is directly related to the concentration of a compound in the sample. However, urine samples are generally arbitrarily diluted, which makes the compound concentrations inappropriate measures for its composition—specifically considering the data-analysis goal of most NMR urine studies. These uninteresting dilution differences therefore need to be removed from the data by pre-processing. "Normalization" does this by estimating the "dilution factor" by which each sample has to be corrected—*e.g.*, by expressing the intensity of each signal relative to the creatinine signal, a normally-occurring metabolite which is assumed to be excreted at a stable rate. Another normalisation method corrects by equalising the integral of each spectrum, although this can introduce spurious correlations among different chemical shifts [33]. Also differences in chromatographic peak heights—of the same analyte—may be circumvented by normalisation [34].

Even if all artefacts are removed from the data, it may not be the case that all measured variables contain information that is important for the scientific question at hand. Moreover, important variables can be masked by the variability of many unimportant variables, specifically—but by far not exclusively—when variables are expressed on different scales. Scaling is used to equalise the *a priori* potential of each variable to contribute to the eventual model. Which type of scaling should be applied partly depends on the model that will be created after pre-processing. For example, PCA describes the largest variation in the data, which means that variables with higher variance are likely to dominate the fitted model. A commonly-used scaling method to reduce this effect is autoscaling (also known as "Unit-Variance" scaling or "standardisation"), where one divides all measured values for a

variable by the standard deviation of these values, after having centred the variable to its mean. However, unimportant variables that only contain noise—and therefore have low standard deviation—become very important upon autoscaling. Other well-known scaling methods include Pareto scaling, level scaling and VAST scaling [3], which each aim to equalise the contribution of different variables by assigning different weights to each variable.

An alternative to scaling can be found in element-wise transformations of the original data. These transformations are different from scaling because they modify individual elements of the data matrix rather than entire variables. Examples of such transformations include the power and log transformation, which simply take the square root or log of each individual data element, respectively. Transformations implicitly correct for heteroscedastic noise [3]. Specific transformations may also reduce the importance of large values (in the case of power and log transformations), bringing about a "pseudo-scaling" effect [3]. Finally, "rank transformations" replace the original values by their ranks, and that dramatically reduces the influence of outliers, but also hampers interpretation of the information in the fitted model.

2.2.6 Supervised pre-processing methods

A somewhat different class of pre-processing methods is found in "orthogonal signal corrections", of which (Direct) Orthogonal Signal Correction [(D)OSC] is the most common [11, 35]. This type of pre-processing assumes that all variation unrelated to a response variable (such as a concentration or class vector) is an artefact and should be filtered. An advantage of these methods is that multiple artefacts (*e.g.*, a baseline slope plus scatter) can be corrected at once. OSC and DOSC aim to remove all variance in the signals unrelated to the response variable(s) based on a Partial Least Squares (PLS)-like decomposition. The foregoing implies that these methods can only be applied when a response variable is present—*i.e.* these methods are *supervised* pre-processing methods. As we show in section 2.3, the presence of a response variable does not automatically deem these methods useful in a pre-processing strategy.

2.2.7 Artefacts in coupled techniques

The pre-processing methods presented in the previous sub-sections were all designed to deal with two-way data matrices (samples by variables). However, many commonly-used analytical techniques combine multiple individual analytical chemical techniques into coupled techniques, which result in far-more complex data. Combinations between chromatography and mass spectrometry (MS)–based detection are widely used (GC-MS,

LC-MS and CE-MS, where CE stands for Capillary Electrophoresis), but there are also combinations with IR detection (GC-IR) and NMR detection (LC-NMR). The same artefacts as in the individual techniques also apply to these data.

However, application of the same pre-processing methods to correct for these artefacts is not always straightforward: the pre-processing method should explicitly take the multi-way data structure into account. For scaling and transformations, this is clarified in [36], where it is shown that applying a certain type of scaling subsequently on two different data dimensions leads to unwanted side effects.

Some pre-processing methods have been specifically developed for pre-processing multi-way data. For instance, PTW and COW are adapted for dealing with LC-MS and GC-MS allowing alignment of two-dimensional data at the same time [37, 38]. An overview of pre-processing tools for LC-MS metabolomics data is available [18]; a more detailed description of the full data-processing strategy for metabolomics data is also available [39]. Many more 2D pre-processing methods exist, but it is outside the scope of this chapter to elaborate more extensively on this subject.

2.3 Approaches to select a pre-processing strategy

Section 2.2 showed that a multitude of methods exist that correct for specific data artefacts. Naturally, corrections should only be made for artefacts that are present in the data, to avoid the introduction of additional, unwanted artefacts or variation to the data. Avoiding this is not always possible, because observing artefacts in the raw data may be extremely difficult, specifically when multiple artefacts are present. Artefacts may also be multivariate in nature—*i.e.* visible only on combinations between variables, but not on individual variables. Moreover, even artefacts known to be present may require pre-processing specifically suited to the dataset (also see sub-section 2.3.1). It is crucial to select the optimal pre-processing strategy for the data set at hand, as pre-processing can make or break subsequent data analysis [40].

Despite the importance of the selection of optimal pre-processing strategies, it appears that selection approaches are not very common. In many cases, pre-processing strategies are selected based on identification of artefacts or on past experience, but this does not guarantee proper selection, as highlighted in the previous paragraph. Based on an extensive literature study, it appears that three types of pre-processing selection approaches are commonly used—if used at all—to choose the optimal pre-processing strategy:

1. Trial and error, in which one applies different pre-processing strategies and selects the best performing one(s) according to the goal of data analysis;
2. Visual inspection, where, for each pre-processing strategy of interest, the pre-processed data is inspected by simply looking at these data and checking if any artefacts are still visible;
3. Assessment of pre-processed data by data quality parameters, which aim at quantifying the presence of artefacts in the data.

The "trial and error" approach is a fit-for-use approach where the model of interest (*e.g.*, a calibration model) is applied to the pre-processed data. Subsequently, the results of these models are used to assess how the applied pre-processing strategy contributed to the data-analysis goal. "Visual inspection" and "assessment by data quality parameters" are both independent of the data-analysis goal. Visual inspection is the simplest approach and targets specific data artefacts. Assessment by quality parameters is based on the decrease or increase of quantitative parameters that assess data quality upon pre-processing.

After a literature review of these three selection approaches, we will critically assess and compare all three strategy-selection approaches, based on a spectroscopic data set. All pre-processing methods included in this study are deemed sensible and physically meaningful for the current data set, and therefore have a rightful place in the data pre-processing strategy.

2.3.1 Data and methods

Throughout this chapter, the properties of the three pre-processing selection approaches are illustrated by application to FTIR data originating from a food authentication study. More specifically, this study considered the authentication of Rochefort beers [41]. One of the goals in this study was to discriminate Rochefort 8° beers from the other Rochefort beers, those being Rochefort 6° and 10°. Specifics regarding sample collection and spectroscopic measurements are described by Engel *et al.* [41].

The raw FTIR data is displayed in figure 2.4, which clearly shows variation between the samples unrelated to their class. This artefact seems to be related to the batch in which the samples were measured. To cover possible seasonal variability, the samples were collected in September 2008 and January 2009.

Based on visual inspection of the raw data, pre-processing should correct for the variation among different sample-measurement batches. However, note that not all artefacts may be visible to the naked eye; so all pre-processing methods that correct artefacts in spectro

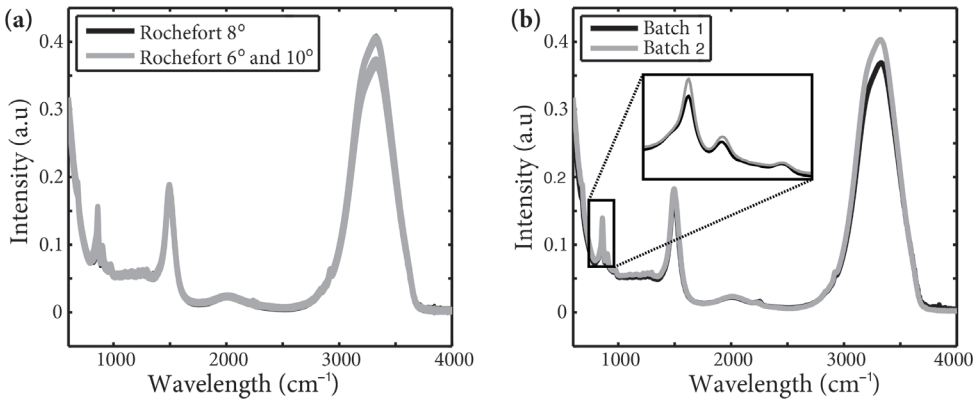

Figure 2.4 **(a)** Raw FTIR data, where Rochefort 8° beers are indicated with a black line and Rochefort 6° beers and Rochefort 10° beers with a grey line. **(b)** Example of two FTIR spectra of beers from one class, measured in different batches. The 'batch effect' is clearly visible.

scopic data (see table 2.1) were selected. Noise reduction, scaling and transformations were also included. Table 2.2 lists all pre-processing methods that were considered. Note that in total 4914 combinations of methods were investigated.

All pre-processed data sets were used as input to a PLS-LDA classifier to predict the beer class [42]. A 10-fold cross-validation was used to determine the optimal number of latent variables (LVs) in the PLS model as well as the percentage of correct classification. The optimal number of LVs will be referred to as the *model complexity* and the percentage of correct classification for this number of LVs as the *classification accuracy*. The accuracy takes the different number of samples that were present in both classes into account (*i.e.* it is a geometric mean instead of an arithmetic mean).

2.3.2 Trial and error approaches

The goal of pre-processing selection via trial and error is to select the best method or combination of methods from a set of likely candidates. All pre-processing approaches are applied to the data and subsequently the pre-processed data is used as an input to a classification or calibration model. This model is used to assess the quality of the pre-processing strategy by an internal measure, such as RMSEP or RMSECV [11, 43, 44].

Commonly, only a small set of candidate pre-processing methods is investigated: combinations of methods, such as a baseline correction followed by scatter correction, are not always taken into account. For example, Wold et al. compared no pre-processing of NIR data with MSC and OSC [11]. In two separate studies on NIR data, Luypaert *et al.* investigated no

Table 2.2 Overview of all pre-processing methods that were applied to the FTIR data. All combinations of methods in steps 1 – 4 were considered, except for supervised pre-processing methods (see section 2.2.6), which involved combinations of OSC and DOSC only with step 4. Note that OSC and DOSC correct for baseline, scatter effects, and noise at the same time.

Step 1: Baseline	Step 2: Scatter	Step 3: Noise	Step 4: Scaling & Transformations
No baseline correction	No scatter correction	No noise removal	Meancentering
Detrending, 2nd order polynomial	Mean scaling	S-G smoothing (window : 5 pt, order: 2)	Autoscaling
Detrending, 3rd order polynomial	Median scaling	S-G smoothing (window : 9 pt, order: 2)	Range scaling
Detrending, 4th order polynomial	Maximum scaling	S-G smoothing (window : 11 pt, order: 2)	Pareto scaling
AsLS	L2 norm scaling	S-G smoothing (window : 5 pt, order: 3)	Poisson scaling
Derivatisation with 1st order polynomial	SNV	S-G smoothing (window : 9 pt, order: 3)	Level scaling
Derivatisation with 2nd order polynomial	RNV, 15%	S-G smoothing (window : 11 pt, order: 3)	Log transformation
	RNV, 25%	S-G smoothing (window : 5 pt, order: 4)	
	RNV, 35%	S-G smoothing (window : 9 pt, order: 4)	
	MSC	S-G smoothing (window : 11 pt, order: 4)	
Supervised pre-processing methods			
OSC			Meancentering
DOSC			Autoscaling
			Range scaling
			Pareto scaling
			Poisson scaling
			Level scaling
			Log transformation

pre-processing, SNV, derivatives and detrending; and no processing, SNV, MSC, derivatives, offset correction, detrending and DOSC respectively [45, 46]. In yet another study on NIR data, Candolfi *et al.* selected no pre-processing, offset correction, detrending, SNV, SNV + detrending, MSC and derivatives for trial and error [47]. When comparing the results of these studies it is evident that no single method is best for pre-processing of NIR data. In these four studies, OSC, SNV + derivatisation, DOSC, and SNV performed best, respectively. In other words, many different methods are suitable for pre-processing of NIR data. However, which method is optimal depends on the data set characteristics and the goal of data analysis. This raises the question which methods should be selected for trial and error. Obviously, the same holds true for data from other analytical techniques.

To assess this selection approach critically, we applied 4914 pre-processing strategies we considered as being "sensible" to a two-class classification problem in FTIR data (see sub-section 2.3.1 for details). The effect of pre-processing on classification accuracy and complexity of the model is shown in figure 2.5. Only 273 out of 4914 strategies resulted in a more accurate, less complex model, compared to the model based on the raw data. This means that in most cases sensible pre-processing strategies actually increased the model complexity or worsened the model accuracy by up to 20%.

Also, it was found that the effect of pre-processing methods depends on other methods within the strategy. For example, many combinations that included smoothing improved the accuracy and complexity, but an equally large number of combinations did not. Therefore, the optimal pre-processing strategy cannot be found by sequentially optimising each artefact, but strategies that combine all methods should be considered in this trial and error approach. This highlights that the holistic overview on data pre-processing, which the trial and error approach provides, may be very computationally intensive.

Search algorithms may considerably reduce this time immensely (*e.g.*, genetic algorithms may be used to optimise the pre-processing strategy) [48-50]. However, Jarvis *et al.* note that their optimisation took five days for FTIR data with 850 spectral variables and 360 samples. Considering that *e.g.*, metabolomics and proteomics platforms provide data with numbers of variables that can be magnitudes higher, trial and error strategy selection through genetic algorithms is at least not generally applicable for all analytical chemical data.

2.3.3 Visual inspection

The computational intensity of trial and error approaches can be greatly reduced by assessment of the effect of pre-processing before a model is constructed. This is the goal of visual

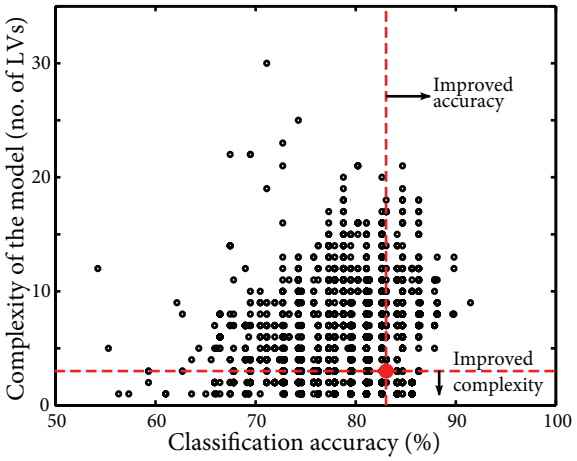

Figure 2.5 Classification accuracy versus complexity of the model for all pre-processing strategies (see subsection 2.3.1). The classification accuracy and complexity for the raw data are given by the large red dot. All strategies to the right of the vertical dotted line improve the classification accuracy compared to the raw data; all strategies to the right of the vertical dotted line and below the horizontal dotted line improve both the classification accuracy and the complexity of the model.

inspection. Samples should show more spectral overlap after pre-processing in visual inspection (*i.e.* artefacts have been removed) and, if applicable, differences between groups of samples should be more pronounced. As noted by Giskeodegard *et al.*, the human eye and brain are still unsurpassed as pattern recognition tools [51], so they are invaluable tools for assessing the quality of pre-processing. At the same time, these tools are not objective: it is, for example, a matter of perception whether differences between groups have become more pronounced or not.

Visual inspection of the pre-processed spectral data is often performed after peak alignment [51]. However, in our experience judging alignment quality in crowded spectra, such as NMR spectra of human urine, remains a challenging task. Visual inspection of spectral data is also common practice in other fields. For example, a reduction in variation between NIR spectra was observed after SNV pre-processing in ref. [52].

Visual inspection of pre-processing may be very difficult and rather subjective for subtle artefacts, so many reports in the literature do not inspect the data in "spectral mode" but in a lower dimensional space [53]. Many methods can be used to reduce the dimension of data, but PCA is by far the most popular option used to judge pre-processing quality as it is unsupervised and therefore 'unbiased' towards the data-analysis goal [54]. In [55], for example, PCA was used to determine optimal normalisation and alignment of

chromatographic spectra. It was observed that between-class distance increased and within-class distance decreased after alignment.

However, clustering in a PCA score plot is not always an indication of good pre-processing. For example, in a metabolomics study, the influence of various centering, scaling and transformation methods on GC-MS data was assessed by PCA [3]. The clear clusters within the score plots of range scaled and centered data identified both as viable pre-processing methods. However, the corresponding loading plots indicated that only range scaling made sense from a biological viewpoint, since the differences between the groups were related to many metabolites (as is the biological expectation in the data under study in [3]) instead of just a few. This shows that the chemical background of the problem is also essential to evaluate the chosen pre-processing strategy.

Unfortunately, even when including loading plots, visual inspection may still lead to inconclusive results. Torgrip *et al.* have shown this for normalisation of urine NMR data [32]. In this study, the samples were diluted by known constants. Subsequently, it was investigated whether a normalisation method (Histogram Matching, HM) could determine these constants more accurately from the data compared to other methods. It was found that the lowest error was indeed obtained with HM. However, PCA score plots did not show any significant improvement of HM estimation compared to other normalisation methods, neither did inspection of the loadings. Based on PCA, the user might have selected another normalisation method as best.

Gabrielsson *et al.* compared the raw data with the pre-processed data matrix via O2PLS (Orthogonal2-PLS) [56]. O2PLS separates the variation in two matrices—raw and pre-processed data in this example—into three contributions: 'joint' variation shared between raw and pre-processed data and variation unique to either raw or pre-processed data. The scores and loadings of the orthogonal variation can be used to assess the pre-processing quality—even whether it introduced artefacts into the raw data. However insightful and robust, visual inspection of such a model for each pre-processing strategy is simply not feasible for so many candidate pre-processing strategies, a drawback of course shared by all visual inspection approaches described here.

The risk of pre-processing selection via PCA is illustrated for the FTIR example discussed in sub-section 2.3.1. We have selected three different pre-processing strategies for visual inspection, which are all highlighted in panel a in figure 2.6. These three strategies include no pre-processing (*i.e.* classification of the meancentered data, panel b), the most accurate strategy (panel c) and a generally poorer-performing strategy (panel d). For all three

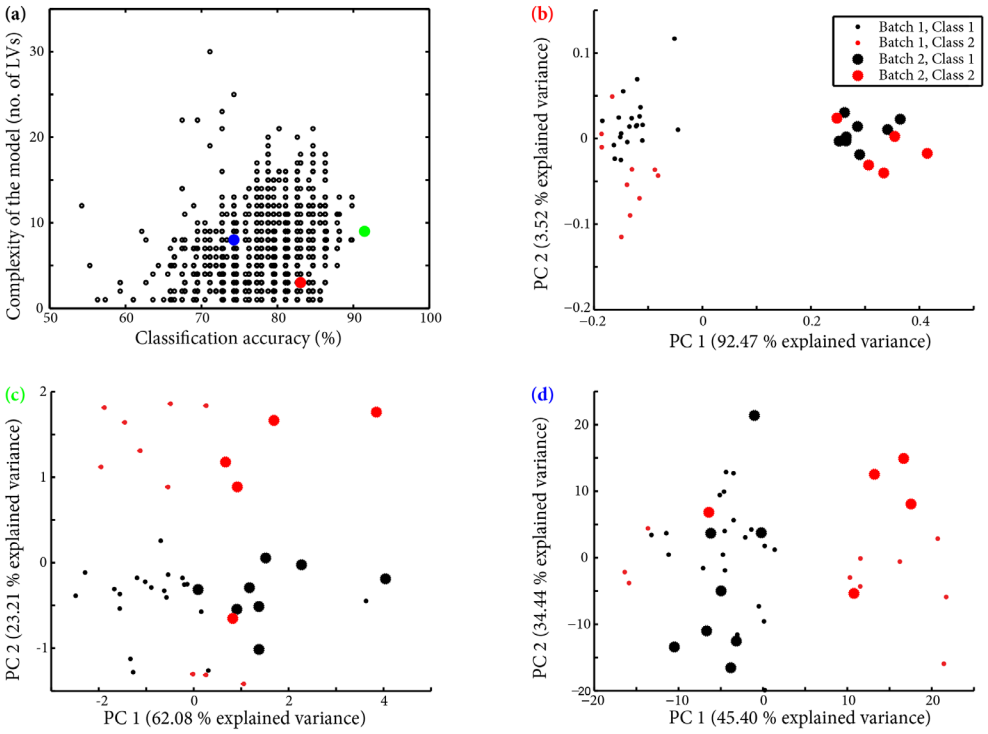

Figure 2.6 Visual comparison of three different pre-processing strategies. **(a)** The accuracy and complexity of the corresponding model for three selected strategies. **(b)** The score plot of the raw data (red dot in panel (a)). **(c)** The score plot of the best pre-processing strategy (green dot), and **(d)** the score plot of a poorer pre-processing strategy (blue dot).

strategies, the score plots of PC 1 versus PC 2 are given in their respective panels. The data contain two batches and two classes, where separating the classes comprises the data analysis goal. Panel b clearly shows that the batch effect is associated to the majority of all variation in the data: PC 1—the direction in which the batch effect is visible—explains over 90% of all variation. However, both classes are also separated reasonably well in the raw data.

Panel c shows that the most favourable pre-processing strategy reduces—but does not remove—the batch effect. This strategy however does not really seem to improve separation between classes, although in a PLS-LDA model it appears to be the most accurate strategy.

The more or less identical between-class separation in panel d suggests a model accuracy similar to that in panel c. Additionally, panel d shows how this strategy completely removes

the batch effect from the data. Although visual inspection thus pinpoints the latter strategy as most appropriate, it has a 17% lower accuracy than the most accurate strategy in a PLS-LDA model and an almost 10% lower accuracy than a model of the raw data. This highlights a major risk of visual inspection of the data compared to trial and error approaches. Visual inspection necessarily filters all visible data artefacts before data analysis, but this does not always lead to optimally accurate models: the batch effect turns out unimportant for the classification accuracy and visual inspection led to a focus on an artefact irrelevant to model accuracy.

2.3.4 Quality parameters

The third approach for pre-processing selection involves the use of quality parameters. These parameters aim to provide an objective measure of the quality of the pre-processed data. Here, data of a higher quality is defined as data with fewer artefacts. The advantage of such methods compared to trial and error approaches is that they do not require computation of calibration or classification models, thereby greatly reducing computational intensity. Quality parameters offer a robust, objective and sometimes quicker alternative to visual inspection. However, pre-processing strategy selection via quality parameters does not yet seem to be common practice.

The 'simplicity value' was introduced by Skov *et al.* to assess the quality of chromatographic alignment [12], and it has since been extended to assess alignment of NMR spectra [13, 51]. To compute the simplicity value, PCA is applied to the pre-processed data. Note that the data is not centered unless required by the pre-processing strategy. Next, the variance explained by the principal components is assessed. The assumption is that correct pre-processing will remove variation from the data. Therefore, the first PCs of correctly pre-processed data will have a higher percentage of explained variance compared to the raw data. The simplicity is given as a number between 0 and 1, with higher values indicating better pre-processing. Although the simplicity value is mainly applied in assessing the quality of alignment, we think that it can be useful in assessing the removal of other artefacts as well.

In addition to the simplicity value, the Pearson correlation coefficient can also provide useful insights into the way to select the best pre-processing strategy. Wu *et al.* showed this for a peak alignment algorithm called 'Fuzzy Warping' on urine NMR spectra [57]. The concept behind the correlation is similar to that behind simplicity: removing data artefacts will make spectra more similar, hence increasing their correlation. The correlation can also be used as a parameter to optimise in alignment, as shown by Skov *et al.* for their COW alignment method [12].

The 'peak factor' quality parameter originates from chromatographic alignment [12, 51]. An appropriate pre-processing strategy should leave the shape of peaks intact, something not taken into account with the simplicity value. Instead, the peak factor studies the normalised differences between pre-processed and raw spectra. The smaller the sum of distances, the smaller the differences; this indicates better pre-processing. Since this parameter relates to peak deformations, which are most common in alignment, this parameter is unsuitable for the general purpose of evaluating the quality of pre-processing. For example, methods such as derivatization (see section 2.2) clearly alter peak shapes.

Esquerre *et al.* discuss some quality parameters for pre-processing of hyperspectral NIR images; all measures are based on PCA analysis of the pre-processed data [58]. Interestingly, they use supervised quality parameters besides the unsupervised methods that have been discussed so far, and that may specifically benefit classification problems since, in such cases, the pre-processing objective is to decrease spectral overlap between classes. More specifically, the separation of the two classes of interest along PC1 is assessed by the coefficient of variation (CV). Note that this measure can be used for two-class problems only. Also, the classes should be separated along only PC1. As shown in figure 2.6, this is not always the case. These issues might be circumvented by using a multivariate coefficient of variation measure [59]. This multivariate CV can also be used for multiclass problems and can be applied to the data directly, meaning that no PCA step is required. However, in our study of the beer data, the multivariate CV parameter did not seem to provide interesting information.

Also, Esquerre *et al.* used the explained variance of PC1 (similar to the simplicity value) and the number of outliers as quality parameters. Good pre-processing should be applicable to the majority of the data, and therefore not introduce many outliers, so this is an extremely interesting quality parameter. However, given the assumption that outliers can already be present in the raw data, it is striking that Esquerre *et al.* used a non-robust PCA algorithm to compute the coefficient of variation and simplicity values.

Most likely, no single quality parameter will be able to capture all benefits and drawbacks of a specific pre-processing strategy, so Esquerre *et al.* combined the results of the three quality parameters into one "super-parameter", which was used to judge the pre-processing strategies. In their application, this seemed highly beneficial.

As shown in figure 2.7, no single quality parameter was able to detect the correct pre-processing strategy for the FTIR beer data set either. Here, only the simplicity value and correlation are shown, but similar results were found for other quality parameters.

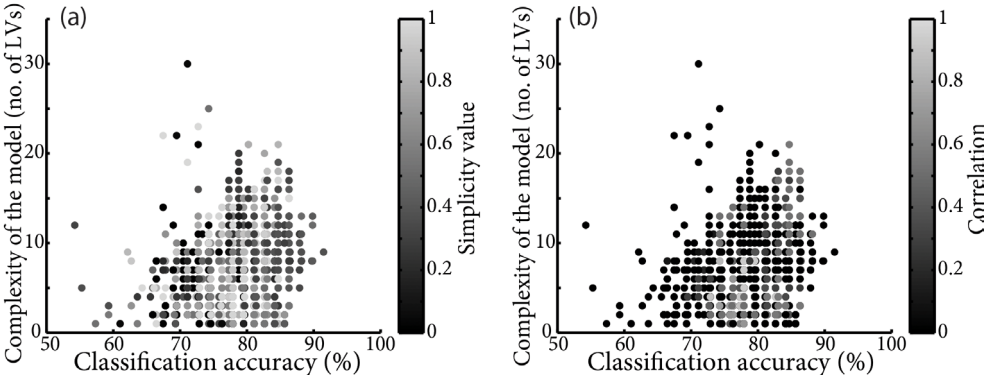

Figure 2.7 The relation between simplicity value **(a)** and correlation **(b)** of differently pre-processed data and the complexity and accuracy of models based on this data.

Construction of a super-parameter was beyond the scope of this review paper, because not much is known on construction of these parameters. Separate studies will be necessary to investigate this subject more extensively. However, the work of Esquerre *et al.* shows the great promise of a quantitative super-parameter.

2.4 Concluding remarks

Data pre-processing—to remove unwanted artefacts and highlight the variation of interest—is crucial to the successful achievement of the data analysis goals, be they e.g., exploration, classification, or prediction. Our example shows how extremely difficult it can be to determine which method—of the vast number of available pre-processing methods—can successfully help to achieve these goals. Also, the specific data set characteristics—*i.e.* which artefacts are present among which other properties of the data—are an additional factor of considerable importance that cannot be ignored in this choice.

All data set properties dictate the combination of appropriate methods that correct for different artefacts. The classification problem presented here led to a more than 20% difference in model accuracy between the best and worst pre-processing strategies. All strategies were reasonable from a theoretical viewpoint. Sub-optimal pre-processing strategies can therefore greatly affect the achievement of the data analysis goals, and that makes a robust and quantitative approach to establishing the optimal pre-processing strategy of the utmost importance.

We have reviewed three widely used approaches (*i.e.* the current trends in pre-processing) for this purpose, and showed that all three have serious drawbacks: they may be time-consuming beyond practicability or may provide misleading results.

The construction of a quality parameter seems very promising, although approaches based on these parameters also provide inconclusive results for our current dataset. The quantitative expression of such a parameter is very intuitive, and the information from different parameters can also be combined into a super-parameter to quantify the effect of pre-processing from different viewpoints simultaneously. These quantitative aspects fit very well into the chemometrics paradigm. In different situations, chemometrics has benefitted very much from such quantitative expressions. For example, it has derived much of its strength from the quantitative power of model validation.

There is an ever-growing body of literature that details and extends such model validation, to the point that a solid theoretical framework has emerged. We envisage a similar theory for data pre-processing—of course fully integrated with that for model validation and for other aspects of chemometrics. This can then be used to establish, or at least predict, the data analysis strategy that combines the methods that optimally remove the artefacts from the data and thereby highlight the information that is required to achieve any of the possible data-analysis goals.

2.5 Pre-processing in this thesis

The rest of this thesis is concerned with analysis of urine ^{1}H-NMR metabolomics data. Common artefacts in such data include baseline drift, peak shifts between samples e.g. due to temperature fluctuations or differences in pH of the samples. Moreover, specific regions of the spectrum contain more peaks. These peaks may overlap. Such crowded regions contain more information but are generally more difficult to analyze as well. Finally, the overall metabolite concentration in urine may vary as a function of the excretion rate and specific physiology of each patient. To allow for sample-to-sample comparison all these artefacts must be removed by data pre-processing; hence the most important pre-processing steps are baseline correction, peak alignment, normalization, and binning.

In this thesis, all spectra were baseline corrected using cubic splines in the Bruker NMR software. Next, icoshift and COW were considered for peak alignment [12, 28]. Creatinine normalization, probabilistic quotient normalization, and histogram matching were used for normalization [32, 60]. Equidistant binning, dynamic adaptive binning and adaptive intelligent binning were considered to the bucket the data [30, 61]. Similar to the

experiment described in sections 2.3.3 and 2.3.4 visual inspection of the data as well as the use of data quality parameters was used to select a suitable pre-processing strategy. It was observed that icoshift introduced artefacts, e.g. artificial horizontal lines or plateaus, to the spectra, especially in crowded regions where a number of peaks overlapped. Therefore, COW was preferred for alignment. The optimal setting was based on the peak factor quality measure. Only small differences were observed between various normalization techniques. Therefore, histogram matching was selected as the most suitable method because other artefacts such as peak misalignments do not influence the results of the method. Adaptive-intelligent binning resulted in better placements of the bins compared to equidistant binning and dynamic adaptive binning. However, all approaches struggled in areas with large peak overlap. Therefore, the result of the AI-binning algorithm was manually adjusted to minimize the risk that a peak was split in two bins.

The optimized pre-processing strategy was used for data pre-treatment in chapter 4. The strategy could also be applied to chapters 5 and 6. These chapters considered the comparison of individual patients to a set of controls. Pre-processing of this data was challenging because it was unknown how the optimized strategy could influence future patient samples; for example, binning or alignment may distort an important peak related to a disease that was not present in the training or test data. Therefore, it was decided to use the conventional pre-processing strategy involving equidistant binning and normalization to creatine in these chapters. This way, it was ensured that the reported results were not overly optimistic. It is clear, however, that further development to select a more appropriate pre-processing strategy for such data is required.

References

1. Eriksson, L., et al., *Using chemometrics for navigating in the large data sets of genomics, proteomics, and metabonomics (gpm).* Analytical and Bioanalytical Chemistry, 2004. **380**(3): p. 419-429.
2. Wold, S., A. Berglund, and N. Kettaneh, *New and old trends in chemometrics. How to deal with the increasing data volumes in R&D&P (research, development and production) - with examples from pharmaceutical research and process modeling.* Journal of Chemometrics, 2002. **16**(8-10): p. 377-386.
3. van den Berg, R.A., et al., *Centering, scaling, and transformations: improving the biological information content of metabolomics data.* Bmc Genomics, 2006. **7**.
4. Rinnan, A., F. van den Berg, and S.B. Engelsen, *Review of the most common pre-processing techniques for near-infrared spectra.* Trac-Trends in Analytical Chemistry, 2009. **28**(10): p. 1201-1222.
5. Parastar, H., M. Jalali-Heravi, and R. Tauler, *Comprehensive two-dimensional gas chromatography (GC x GC) retention time shift correction and modeling using bilinear peak alignment, correlation optimized shifting and multivariate curve resolution.* Chemometrics and Intelligent Laboratory Systems, 2012. **117**: p. 80-91.
6. Forshed, J., I. Schuppe-Koistinen, and S.P. Jacobsson, *Peak alignment of NMR signals by means of a genetic algorithm.* Analytica Chimica Acta, 2003. **487**(2): p. 189-199.
7. Ding, T., et al., *Far-infrared spectroscopy analysis of linear and cyclic peptides, and lysozyme.* Vibrational Spectroscopy, 2012. **61**: p. 144-150.
8. Szymanska, E., et al., *Evaluation of different warping methods for the analysis of CE profiles of urinary nucleosides.* Electrophoresis, 2007. **28**(16): p. 2861-2873.
9. Brown, M., et al., *A metabolome pipeline: from concept to data to knowledge.* Metabolomics, 2005. **1**(1): p. 39-51.
10. Barnes, R.J., M.S. Dhanoa, and S.J. Lister, *Standard Normal Variate Transformation and De-Trending of near-Infrared Diffuse Reflectance Spectra.* Applied Spectroscopy, 1989. **43**(5): p. 772-777.
11. Wold, S., et al., *Orthogonal signal correction of near-infrared spectra.* Chemometrics and Intelligent Laboratory Systems, 1998. **44**(1-2): p. 175-185.
12. Skov, T., et al., *Automated alignment of chromatographic data.* Journal of Chemometrics, 2006. **20**(11-12): p. 484-497.
13. MacKinnon, N., et al., *Variable Reference Alignment: An Improved Peak Alignment Protocol for NMR Spectral Data with Large Intersample Variation.* Analytical Chemistry, 2012. **84**(12): p. 5372-5379.
14. Cen, H.Y. and Y. He, *Theory and application of near infrared reflectance spectroscopy in determination of food quality.* Trends in Food Science & Technology, 2007. **18**(2): p. 72-83.
15. Zeaiter, M., J.M. Roger, and V. Bellon-Maurel, *Robustness of models developed by multivariate calibration. Part II: The influence of pre-processing methods.* Trac-Trends in Analytical Chemistry, 2005. **24**(5): p. 437-445.
16. Daszykowski, M., et al., *Start-to-end processing of two-dimensional gel electrophoretic images.* Journal of Chromatography A, 2007. **1158**(1–2): p. 306-317.
17. Smolinska, A., et al., *NMR and pattern recognition methods in metabolomics: From data acquisition to biomarker discovery: A review.* Analytica Chimica Acta, 2012. **750**: p. 82-97.
18. Castillo, S., et al., *Algorithms and tools for the preprocessing of LC-MS metabolomics data.* Chemometrics and Intelligent Laboratory Systems, 2011. **108**(1): p. 23-32.
19. Brereton, R.G., *Introduction to multivariate calibration in analytical chemistry.* Analyst, 2000. **125**(11): p. 2125-2154.
20. Savitzky, A. and M.J.E. Golay, *Smoothing + Differentiation of Data by Simplified Least Squares Procedures.* Analytical Chemistry, 1964. **36**(8): p. 1627-&.
21. Barclay, V.J., R.F. Bonner, and I.P. Hamilton, *Application of wavelet transforms to experimental spectra: Smoothing, denoising, and data set compression.* Analytical Chemistry, 1997. **69**(1): p. 78-90.
22. Eilers, P.H.C., *Parametric time warping.* Analytical Chemistry, 2004. **76**(2): p. 404-411.
23. Guo, Q., W. Wu, and D.L. Massart, *The robust normal variate transform for pattern recognition with near-infrared data.* Analytica Chimica Acta, 1999. **382**(1-2): p. 87-103.
24. Geladi, P., D. Macdougall, and H. Martens, *Linearization and Scatter-Correction for near-Infrared Reflectance Spectra of Meat.* Applied Spectroscopy, 1985. **39**(3): p. 491-500.
25. Fearn, T., et al., *On the geometry of SNV and MSC.* Chemometrics and Intelligent Laboratory Systems, 2009. **96**(1): p. 22-26.

26. Torgrip, R.O., E. Alm, and K.M. Åberg, *Warping and alignment technologies for inter-sample feature correspondence in 1D H-NMR, chromatography-, and capillary electrophoresis-mass spectrometry data.* Bioanalytical Reviews, 2010. **1**(2-4): p. 105-116.
27. Bloemberg, T.G., et al., *Warping Methods for Spectroscopic and Chromatographic Signal Alignment: A Tutorial Review.* Analytica Chimica Acta, 2013. **(in press)**.
28. Savorani, F., G. Tomasi, and S.B. Engelsen, *icoshift: A versatile tool for the rapid alignment of 1D NMR spectra.* Journal of Magnetic Resonance, 2010. **202**(2): p. 190-202.
29. Tomasi, G., F. Savorani, and S.B. Engelsen, *icoshift: An effective tool for the alignment of chromatographic data.* Journal of Chromatography A, 2011. **1218**(43): p. 7832-7840.
30. De Meyer, T., et al., *NMR-based characterization of metabolic alterations in hypertension using an adaptive, intelligent binning algorithm.* Analytical Chemistry, 2008. **80**(10): p. 3783-3790.
31. Amigo, J.M., T. Skov, and R. Bro, *ChroMATHography: Solving Chromatographic Issues with Mathematical Models and Intuitive Graphics.* Chemical Reviews (Washington, DC, United States), 2010. **110**(8): p. 4582-4605.
32. Torgrip, R.J.O., et al., *A note on normalization of biofluid 1D H-1-NMR data.* Metabolomics, 2008. **4**(2): p. 114-121.
33. Aruga, R., *Closure of analytical chemical data and multivariate classification.* Talanta, 1998. **47**(4): p. 1053-1061.
34. Kultima, K., et al., *Development and Evaluation of Normalization Methods for Label-free Relative Quantification of Endogenous Peptides.* Molecular & Cellular Proteomics, 2009. **8**(10): p. 2285-2295.
35. Westerhuis, J.A., S. de Jong, and A.K. Smilde, *Direct orthogonal signal correction.* Chemometrics and Intelligent Laboratory Systems, 2001. **56**(1): p. 13-25.
36. Bro, R. and A.K. Smilde, *Centering and scaling in component analysis.* Journal of Chemometrics, 2003. **17**(1): p. 16-33.
37. Christin, C., et al., *Optimized time alignment algorithm for LC-MS data: Correlation optimized warping using component detection algorithm-selected mass chromatograms.* Analytical Chemistry, 2008. **80**(18): p. 7012-7021.
38. Bloemberg, T.G., et al., *Improved parametric time warping for proteomics.* Chemometrics and Intelligent Laboratory Systems, 2010. **104**(1): p. 65-74.
39. Hendriks, M.M.W.B., et al., *Data-processing strategies for metabolomics studies.* Trac-Trends in Analytical Chemistry, 2011. **30**(10): p. 1685-1698.
40. Famili, F., et al., *Data pre-processing and intelligent data analysis.* International Journal on Intelligent Data Analysis, 1997. **1**(1).
41. Engel, J., et al., *Confirmation of brand identity of a Trappist beer by mid-infrared spectroscopy coupled with multivariate data analysis.* Talanta, 2012. **99**: p. 426-432.
42. Barker, M. and W. Rayens, *Partial least squares for discrimination.* Journal of Chemometrics, 2003. **17**(3): p. 166-173.
43. Denoord, O.E., *The Influence of Data Preprocessing on the Robustness and Parsimony of Multivariate Calibration Models.* Chemometrics and Intelligent Laboratory Systems, 1994. **23**(1): p. 65-70.
44. Renard, B.Y., et al., *When less can yield more - Computational preprocessing of MS/MS spectra for peptide identification.* Proteomics, 2009. **9**(21): p. 4978-4984.
45. Luypaert, J., M.H. Zhang, and D.L. Massart, *Feasibility study for the use of near infrared spectroscopy in the qualitative and quantitative analysis of green tea, Camellia sinensis (L.).* Analytica Chimica Acta, 2003. **478**(2): p. 303-312.
46. Luypaert, J., et al., *An evaluation of direct orthogonal signal correction and other preprocessing methods for the classification of clinical study lots of a dermatological cream.* Journal of Pharmaceutical and Biomedical Analysis, 2002. **30**(3): p. 453-466.
47. Candolfi, A., et al., *The influence of data pre-processing in the pattern recognition of excipients near-infrared spectra.* Journal of Pharmaceutical and Biomedical Analysis, 1999. **21**(1): p. 115-132.
48. Jarvis, R.M. and R. Goodacre, *Genetic algorithm optimization for pre-processing and variable selection of spectroscopic data.* Bioinformatics, 2005. **21**(7): p. 860 - 868.
49. Bocklitz, T., et al., *How to pre-process Raman spectra for reliable and stable models?* Analytica Chimica Acta, 2011. **704**(1-2): p. 47-56.
50. Laxalde, J., et al., *Characterisation of heavy oils using near-infrared spectroscopy: Optimisation of pre-processing methods and variable selection.* Analytica Chimica Acta, 2011. **705**(1-2): p. 227-234.
51. Giskeodegard, G.F., et al., *Alignment of high resolution magic angle spinning magnetic resonance spectra using warping methods.* Analytica Chimica Acta, 2010. **683**(1): p. 1-11.

52. Luypaert, J., et al., *The effect of preprocessing methods in reducing interfering variability from near-infrared measurements of creams.* Journal of Pharmaceutical and Biomedical Analysis, 2004. **36**(3): p. 495-503.
53. Andersson, F.O., R. Kaiser, and SP. Jacobsson, *Data preprocessing by wavelets and genetic algorithms for enhanced multivariate analysis of LC peptide mapping.* Journal of Pharmaceutical and Biomedical Analysis, 2004. **34**(3): p. 531-541.
54. Sanchez, F.C., P.J. Lewi, and D.L. Massart, *Effect of Different Preprocessing Methods for Principal Component Analysis Applied to the Composition of Mixtures - Detection of Impurities in Hplc-Dad.* Chemometrics and Intelligent Laboratory Systems, 1994. **25**(2): p. 157-177.
55. Hendriks, M.M.W.B., et al., *Preprocessing and exploratory analysis of chromatographic profiles of plant extracts.* Analytica Chimica Acta, 2005. **545**(1): p. 53-64.
56. Gabrielsson, J., et al., *OPLS methodology for analysis of pre-processing effects on spectroscopic data.* Chemometrics and Intelligent Laboratory Systems, 2006. **84**(1-2): p. 153-158.
57. Wu, W., et al., *Peak alignment of urine NMR spectra using fuzzy warping.* Journal of Chemical Information and Modeling, 2006. **46**(2): p. 863-875.
58. Esquerre, C., et al., *Suppressing sample morphology effects in near infrared spectral imaging using chemometric data pre-treatments.* Chemometrics and Intelligent Laboratory Systems, 2012. **117**: p. 129-137.
59. Albert, A. and L.X. Zhang, *A novel definition of the multivariate coefficient of variation.* Biometrical Journal, 2010. **52**(5): p. 667-675.
60. Dieterle, F., et al., *Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics.* Analytical chemistry, 2006. **78**(13): p. 4281-4290.
61. Anderson, P.E., et al., *Dynamic adaptive binning: an improved quantification technique for NMR spectroscopic data.* Metabolomics, 2011. **7**(2): p. 179-190.

3

REGULARIZATION

High dimensional data poses significant challenges for statistical modeling. Many techniques are not applicable to such data, or their results are unreliable. Regularization is a common approach to obtain sensible statistical models for high dimensional data. Most approaches try to reduce the variance of the model at the cost of some bias, i.e. by adding constraints or penalties to the model.

In this chapter we review general regularization approaches in the context of SPC, LDA, and MANOVA. These statistical methods are of main interest in this thesis (see chapter 1). Note that the regularization approaches described in this chapter are general and can also be used for other chemometric methods.

The covariance matrix plays an important role in these methods. Therefore, we first discuss regularization of this matrix. Next, constraints to achieve variable selection are discussed. Applications of these types of regularization in SPC, LDA and MANOVA are provided at the end of the chapter. Finally, we place these methods in the context of the regularized methods that are developed in chapters 4 – 6 of this thesis.

3.1 Introduction

Many experiments in physical and biomedical sciences give rise to an ever-increasing amount of measurement data, originating from multiple advanced analytical technologies [1-4]. Sometimes this is referred to as the data tsunami [5]. The availability of such high dimensional data has reshaped statistical thinking and data analysis [1, 3, 4, 6, 7]. One of the advantages of massive data is that experiments can be performed to generate hypothesis regarding the system under study instead of experiments that aim to prove a hypothesis [8].

In metabolomics, for example, techniques such as NMR and LC-MS are used to measure the concentration of a large number of metabolites in an untargeted way in different groups of samples. Subsequently, data analysis is used to determine if these groups differ significantly and which metabolites are related to the difference. The identified metabolites can be further studied in new, possibly targeted, experiments. This approach has proven successful in many applications such as to find biomarkers for a specific disease [9-12].

Although it is nowadays possible to generate a wealth of data per sample, many statistical methods that perform well for low dimensional data are facing significant challenges in analyzing such high dimensional data [1, 2, 4, 6]. Often these challenges are combined under the term "curse of dimensionality" and include amongst others inaccurate parameter estimation, noise accumulation, and spurious correlation. More details can be found in [4, 6, 13]. For example, a result of the curse of dimensionality is that the precision matrix, that is, the inverse of the covariance matrix, cannot be computed when the number of samples in a data set is lower (n) then the number of variables (p) [14, 15]. Because of this, many techniques such as statistical process control (SPC), linear discriminant analysis (LDA), and multivariate analysis of variance (MANOVA) are not directly applicable to high dimensional data.

Regularization is a common approach to obtain sensible statistical models for high dimensional data [6, 14-17]. Regularization essentially means that some structure is imposed on the model with the aim to reduce its variance at the cost of some bias. For example, Bickel et al showed that the standard LDA model has poor performance when applied to high dimensional data[18]. They showed that much better performance was obtained when the prior assumption that all variables were uncorrelated was included in the model. Clearly this assumption is not always realistic (i.e. in metabolomics) and the resulting model was more biased compared to LDA. However, at the same time its variance was greatly reduced. Because of this, the model generalized better to new data explaining the improved performance.

In this chapter we review regularization approaches for SPC, LDA, and MANOVA. These statistical methods are of main interest in this thesis (see chapter 1). We describe general regularization methods for any of these techniques; most regularization approaches could also be used for other chemometric methods. To be able to describe the regularization methods in a general fashion we first note that the Mahalanobis distance (MD) plays an important role in SPC, LDA, and MANOVA (see section 3.2). Regularization of the three methods essentially comes down to regularization of MDs. Additionally; regularization methods that have been used for any of these techniques can most likely also be applied to the others due to their close connection.

In section 3.2 we study some aspects of the MD in high dimensional data. Based on these considerations we observe two interesting strategies for regularization of the MD, namely (1) using a regularized estimate of the covariance matrix (note that this includes dimension reduction approaches), or (2) by reducing the number of features in the data set by variable selection. These strategies are often used in literature. In section 3.3 the so-called shrinkage approach to regularization of the covariance matrix is described. Section 3.4 deals with variable selection, mainly focusing on ℓ_1-norm constrained models. In section 3.5 examples from the literature on regularization in SPC, LDA, and MANOVA using strategy 1 or 2 are presented. At the end of the present chapter we give a brief overview of the regularization methods for SPC and MANOVA that are used in later chapters in this thesis for analysis of metabolomics data.

3.2 The role of the Mahalanobis distance in chemometrics

Many chemometric techniques essentially involve the measurement of distances [6, 19, 20]. In this section we show that the squared MD plays an important role in SPC, LDA, and MANOVA. Next, we review some properties of the MD in high dimensional data.

3.2.1 The Mahalanobis distance

The squared Mahalanobis distance is essentially a measure of the distance between an object ($\mathbf{x}$) and a group of samples with multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ [19]:

$$MD^2(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})^T \tag{3.1}$$

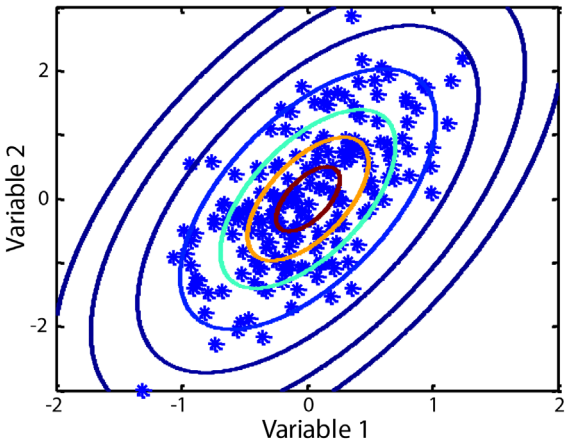

Figure 3.1 Plot of simulated data (blue stars) for two variables. The ellipses represent equal MDs towards the centre of the data points (asterixes).

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ indicate the mean and covariance matrix of $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, respectively. Details such as equations for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ can be found in numerous references such as [6, 19-21]. A general discussion of estimation of $\boldsymbol{\Sigma}$ is presented in section 3.3. The MD can be interpreted as a multivariate generalization of measuring how many standard deviations (σ) object x is away from the centre of the group μ, i.e. $(x-\mu)^2/\sigma^2$. An example is shown in figure 3.1. The ellipses indicate points with equal distance from the group centre.

As will be shown below, many techniques estimate the MD by first finding the linear combination of the variables that maximizes the following Rayleigh criterion [20]:

$$\hat{\mathbf{a}} = argmax_{\mathbf{a}}\left(\frac{\mathbf{a}(\mathbf{x}-\boldsymbol{\mu})^T(\mathbf{x}-\boldsymbol{\mu})\mathbf{a}^T}{\mathbf{a}\boldsymbol{\Sigma}\mathbf{a}^T}\right) = \frac{([\mathbf{x}-\boldsymbol{\mu}]\boldsymbol{a}^T)^2}{\mathbf{a}\boldsymbol{\Sigma}\mathbf{a}^T} \tag{3.2}$$

Geometrically, $\mathbf{a}$ (commonly referred to as a canonical variate) can be interpreted as the direction in the data along which the sample differs most from the centre of the group relative to the variance within the group. An example is shown in the first row of figure 3.2. It can be shown that the solution to (3.2) equals $\mathbf{a} = (\mathbf{x}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}$ [20]. Due to the definition of $\mathbf{a}$, the squared MD can be computed by considering the scalar projections of the data on $\mathbf{a}$, i.e. $MD^2 = ([\mathbf{x}-\boldsymbol{\mu}]\boldsymbol{a}^T)^2/(\mathbf{a}\boldsymbol{\Sigma}\mathbf{a}^T)$. Often, the coefficients of $\mathbf{a}$ are studied to determine on which variables the sample differed most from the group, where high absolute coefficients are assumed to indicate large differences.

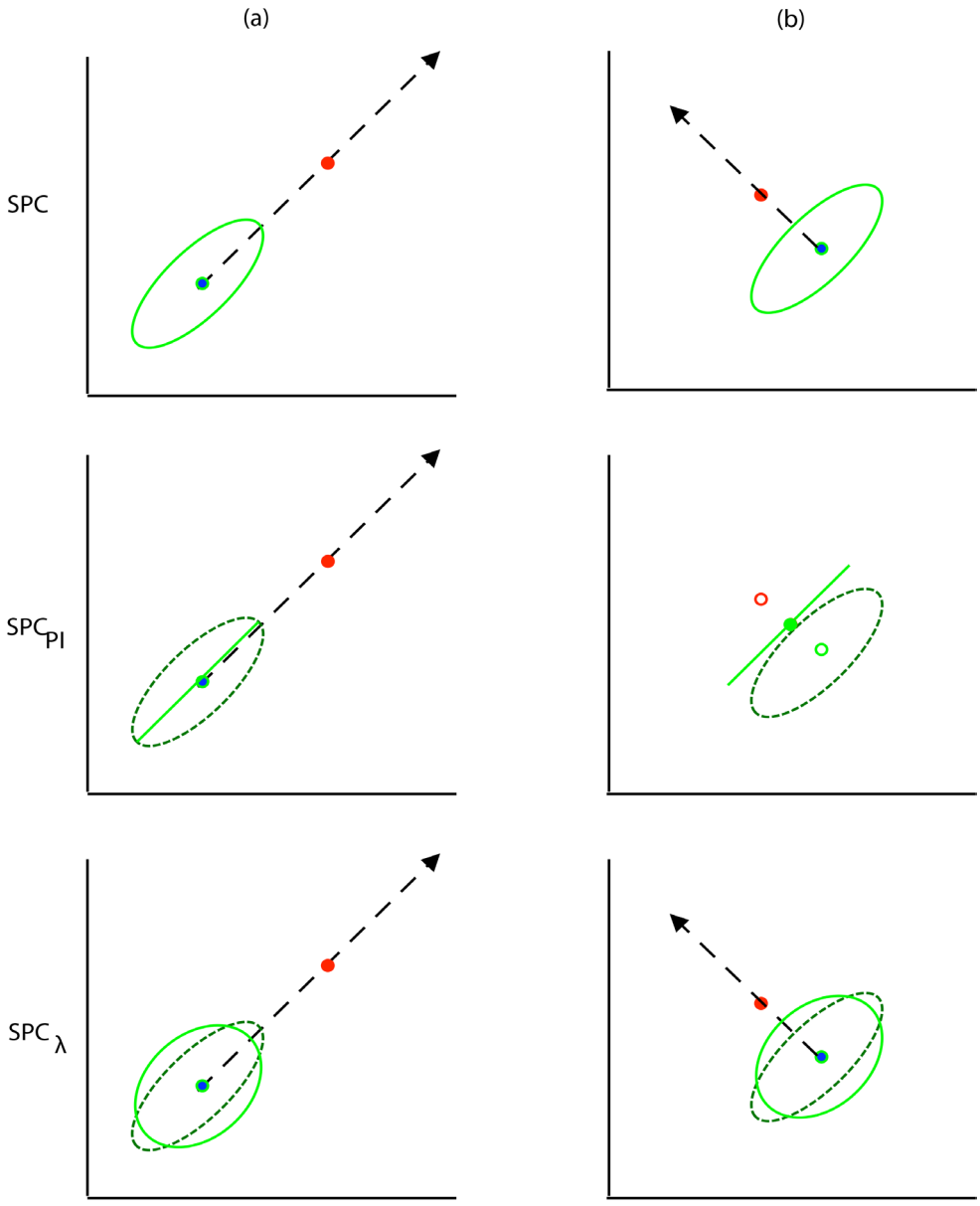

Figure 3.2 Schematic representation illustrating in two dimensions (regularized) SPC methods. Two situations are considered: the outlier (red sample) differs from the controls (green sphere) in the direction of **(a)** the first eigenvector, or **(b)** the second eigenvector of $\mathbf{\Sigma}_C$. The discriminant direction $\mathbf{a}_{SPC}$ is indicated by a dotted line. The first row corresponds to the use of the SPC model defined in equation 3.3 the second row to using equation 3.3 with a pseudo-inverse of $\mathbf{\Sigma}_C$, and the third row to using equation 3.3 with a shrinkage estimator of $\mathbf{\Sigma}_C$. For clarity, the estimated control sphere of the standard SPC model (row 1) is indicated by the dark green dotted sphere in the regularized models (row 2 and 3).

Note that equation 3.1 assumes a multivariate normal distribution. Metabolomics data are often not normally distributed [22]. For example, a normal distribution allows for negative concentrations, which is not very realistic. A log-normal distribution might give a more accurate description of the data. However, for the data in this thesis we have not observed improved results when analyzing log-transformed data. Therefore, we will assume a multivariate Normal distribution in the rest of this chapter.

3.2.2 Statistical process control

The aim of statistical process control (SPC) techniques is to determine if a sample ($\mathbf{x}_p$) differs significantly from a group of control samples ($\mathbf{X}_c$) [23, 24]. An example is the comparison between a patient and a set of healthy controls as we will consider in chapters 5 and 6. Note that SPC is closely related to outlier detection techniques [23, 25].

Often it is assumed that the control group follows a multivariate normal distribution $N(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$, where $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ indicate the mean and covariance matrix of the control samples, respectively. Typically, the null hypothesis of no significant difference between $\mathbf{x}_p$ and $\mathbf{X}_c$ is rejected if [26]:

$$MD^2(\mathbf{x}_p) = (\mathbf{x}_p - \boldsymbol{\mu}_c)\boldsymbol{\Sigma}_c^{-1}(\mathbf{x}_p - \boldsymbol{\mu}_c)^T > c \tag{3.3}$$

This is clearly a squared MD. When the sample estimates for the mean and covariance matrix are used, the upper limit c is obtained from a scaled F-distribution [23]. Geometrically, 3.3 tests whether the sample of interest falls inside the confidence sphere of the control samples, i.e. an ellipse in the first row of figure 3.2. Note that expression 3.2 can be used to determine on which variables $\mathbf{x}_p$ differed from the controls. This information can be used in a subsequent step to investigate the cause of the difference.

3.2.3 Linear discriminant analysis

The goal of LDA is to compare a sample ($\mathbf{x}_p$) to multiple groups or classes and predict to which class it belongs, e.g. to predict whether a sample is healthy or diseased [6, 20, 21]. The method assumes that the classes are multivariate normal with different means but the same covariance matrix. Note that new samples are always assigned to one of the classes specified in the model. The maximum likelihood LDA rule assigns $\mathbf{x}_p$ to the class i that minimizes:

$$MD_i^2(\mathbf{x}_p) = (\mathbf{x}_p - \boldsymbol{\mu}_i)\boldsymbol{\Sigma}_W^{-1}(\mathbf{x}_p - \boldsymbol{\mu}_i)^T \tag{3.4}$$

where $\boldsymbol{\mu}_i$ is the mean of the ith class and $\boldsymbol{\Sigma}_W$ is the covariance matrix that describes the variation of the samples within each class. Geometrically, LDA assigns a sample to the class towards which it has the smallest squared MD. When the number of classes is two the rule allocates $\mathbf{x}_p$ to the first class when:

$$(\mathbf{x}_p - \boldsymbol{\mu})\mathbf{a}_{LDA}^T > 0 \tag{3.5}$$

where $\mathbf{a}_{LDA} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\boldsymbol{\Sigma}_W^{-1}$ and $\boldsymbol{\mu} = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$ [20]. Note that direction $\mathbf{a}_{LDA}$ is closely related to the direction $\mathbf{a}$ estimated by expression 3.2. However, this direction maximizes the difference between the two classes of interest instead of the difference between the sample and one group.

Fishers discriminant rule is closely related to the maximum likelihood LDA rule. They give the same solution for when dealing with two classes [20]. In Fishers approach to LDA, direction $\mathbf{a}_{LDA}$ is first estimated. Subsequently, allocation of a new sample can be done using a distance measure based on the scalar projections of the data onto direction $\mathbf{a}_{LDA}^T$ i.e. using expression 3.5. When dealing with more than 2 classes a single direction is not enough to correctly predict class membership of a new sample. In this case multiple directions are used. These directions are essentially found by maximizing the following expression (with the added constraint that the directions are orthogonal to each other) [20]:

$$\hat{\mathbf{a}}_{FLDA} = argmax_{\mathbf{a}} \frac{\mathbf{a}^T\mathbf{S}_B\mathbf{a}}{\mathbf{a}^T\mathbf{S}_W\mathbf{a}} \tag{3.6}$$

Matrices $\mathbf{S}_B$ and $\mathbf{S}_W$ are so-called scatter matrices. They have been traditionally used in Fisher LDA and are closely related to covariance matrices[1]. Actually, the same solution is obtained when (3.6) is expressed in terms of covariance matrices. Matrix $\mathbf{S}_B$ describes the scatter (covariance) between the class centroids and matrix $\mathbf{S}_W$ describes the scatter within each class. It can be shown that the solutions of (3.6) are the eigenvectors of $\mathbf{S}_W^{-1}\mathbf{S}_B$ corresponding to the highest eigenvalues [20]. The first eigenvector (with the highest eigenvalue) is the direction along which the classes are most separated from each other; the second eigenvector is a direction orthogonal to the first with the second most separation; etc. This shows that expression 3.6 can be interpreted as an extension of (3.2) when dealing with more groups. Typically $g - 1$ directions are inspected when dealing with g groups.

[1] Note that covariance matrices and scatter matrices are closely related. For example here $\boldsymbol{\Sigma}_W = \mathbf{S}_W/(n - g)$, where n is the total number of samples in the data and g is the number of classes. Note that an estimate of a scatter matrix is essentially given by its sample covariance estimate times the degrees of freedom.

Allocation of a new sample can be done by projecting the data onto the $g-1$ directions and using a distance measure such as the squared MD in this lower dimensional space.

3.2.4 Multivariate analysis of variance

MANOVA is mathematically closely related to Fisher LDA (and the MD). However, the method is not used to classify a new sample, but rather to determine if the classes of interest differ significantly [20, 27]. Therefore, it is sensible to apply MANOVA before LDA. Traditionally, MANOVA has been used to analyze experiments that were carried out according to an experimental design, i.e. to test significance of main effects and interactions.

As an example we consider a one-way MANOVA experiment, i.e. one experimental factor was varied at g levels such as healthy, disease stage 1, disease stage 2, etc. It is assumed that all groups are multivariate normal with different means but the same covariance structure. The goal of MANOVA is to assess whether the group means differ significantly, i.e.

$$H_0\text{: } \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_g \text{ vs. } A\text{: at least one mean is different}$$

MANOVA tests are based on the matrices $\mathbf{S}_B$ and $\mathbf{S}_W$. Note that these were defined in expression 3.6 for Fisher LDA, i.e. matrix $\mathbf{S}_B$ describes the scatter between the group centroids and matrix $\mathbf{S}_w$ describes the scatter within the groups. In analogy with ANOVA, a ratio is formed between the between-group scatter ($\mathbf{S}_B$) and the within-group scatter ($\mathbf{S}_W$). More specifically, hypothesis tests are based on the matrix:

$$\mathbf{R} = \mathbf{S}_W^{-1}\boldsymbol{S}_B \tag{3.7}$$

As described for the case of Fisher LDA above, the $g-1$ eigenvectors of $\mathbf{R}$ with the highest eigenvalues are the directions in the data along which the distance between the class means is the largest compared to the variation within the groups, i.e. they are the solution to (3.6). The corresponding eigenvalues are a measure of this distance. The MANOVA significance tests (i.e. Wilks lambda) are based on these eigenvalues. It can be shown that the test-statistic is equal to the MD between the group means when comparing two groups. In other words, in this case MANOVA tests if the squared MD between the group centroids is significantly greater than 0.

3.2.5 What goes wrong in high dimensions?

Traditionally, the MD has been applied to data where the number of variables is limited. The behavior of the MD is, however, very different in data with many variables [4, 13, 19]. Often

these phenomena are not taken into account during data analysis. Below we present three problems that are often observed when the MD is applied to high-dimensional data. The first problem is related to estimation of the sample covariance matrix; the other two problems are related to application of distance measures in high dimensional data.

The sample covariance matrix

The covariance matrix, or more exactly its inverse, is a central aspect of the MD (see, for example, equations 3.1 and 3.7) [19, 20]. Here, we describe the properties of the matrix in the context of process control (equation 3.3). The same principles apply for the other techniques that use the covariance matrix such as LDA and MANOVA.

The unbiased sample estimate of covariance matrix $\mathbf{\Sigma}_c$ is given by:

$$\widehat{\mathbf{\Sigma}}_c = \frac{1}{n_c - 1}\sum_{i=1}^{n_c}\left(\mathbf{x}_{c,i} - \bar{\mathbf{x}}_c\right)^T\left(\mathbf{x}_{c,i} - \bar{\mathbf{x}}_c\right) = \frac{1}{n_c - 1}\overline{\mathbf{X}}_c^T\overline{\mathbf{X}}_c \tag{3.8}$$

where n_c is the number rows in matrix $\mathbf{X}_c$ (number of samples), $\mathbf{x}_{c,i}$ is the ith row in this matrix, $\bar{\mathbf{x}}_c$ is the average over the columns of $\mathbf{X}_c$ (i.e. an estimator for $\boldsymbol{\mu}_c$), and $\overline{\mathbf{X}}_c$ is the data matrix centered with $\bar{\mathbf{x}}_c$ [14, 15, 20]. The advantages of the sample covariance (and corresponding precision) estimate are its ease of computation and the property of being unbiased. However, when the number of samples is smaller than the number of variables the sample covariance estimate cannot be considered a good approximation of the true covariance matrix [14, 15]. This is also true for cases where n_c is close to p in size. In these situations the matrix is potentially estimated with large error. For example, consider the eigenvalues of $\widehat{\mathbf{\Sigma}}_c$ [14]. In figure 3.3, boxplots of the ordered eigenvalues are shown for the case where $\mathbf{\Sigma}_c$ is equal to the identity matrix, i.e. the population eigenvalues are all equal to one. The sample eigenvalues were calculated for simulated data where the underlying distribution was multivariate normal with 100 variables and various numbers of samples. In the figure it can be seen that the largest sample eigenvalues are overestimated while the smallest eigenvalues are underestimated. This effect becomes more severe when the ratio p/n_c becomes larger. When the ratio p/n_c is less than one but not negligible, the sample estimate is numerically ill-conditioned, i.e. inverting it will amplify estimation error [14]. When $p/n_c \geq 1$ matrix $\widehat{\mathbf{\Sigma}}_c$ loses full rank. This can be seen in figure 3.3 as a growing number of eigenvalues become zero. This has several undesirable consequences. First, $\widehat{\mathbf{\Sigma}}_c$ is not positive definite anymore, second it cannot be inverted as $\widehat{\mathbf{\Sigma}}_c$ becomes singular [14].

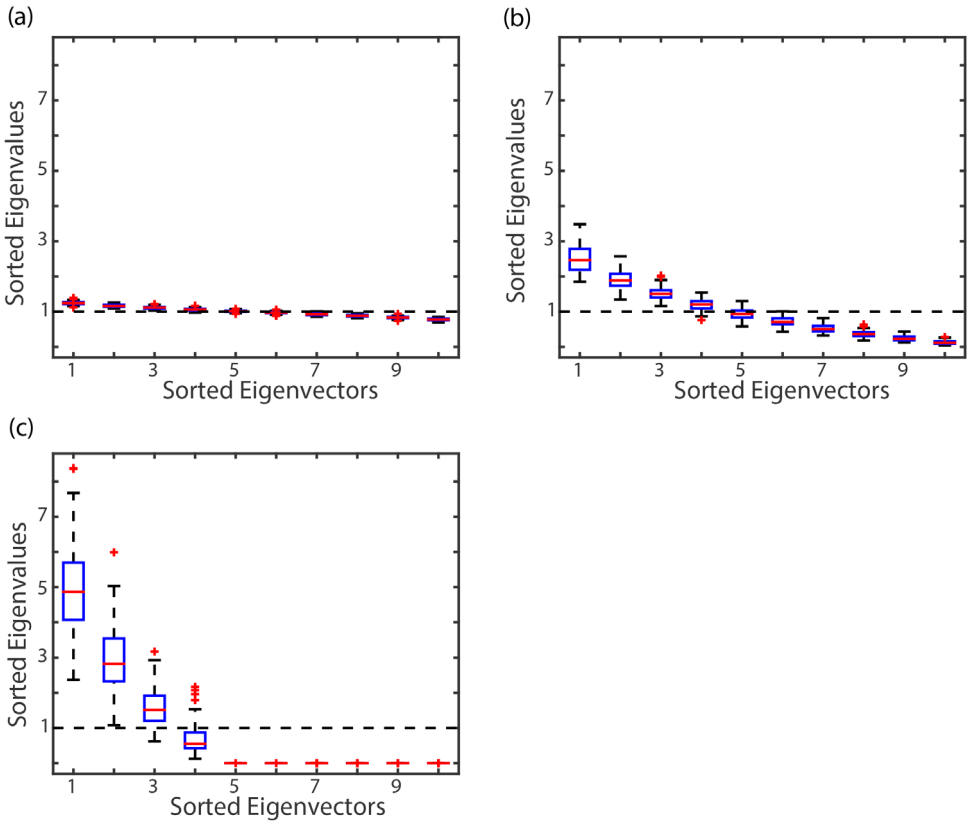

Figure 3.3 The sorted eigenvalues of the sample covariance matrix estimated from 100 simulated data sets. The population covariance matrix was equal to the identity matrix. The population eigenvalues are indicated by the dotted line. The number of variables in the simulation was 10, and the number of samples was **(a)** 500, **(b)** 20, and **(c)** 5.

The bias of the eigenvalues has a negative impact on many techniques, i.e. estimates of explained variance in PCA or estimates of within-group variance in SPC, LDA, and MANOVA. For example, in the context of SPC, the bias of the eigenvalues means that the variance of the control group is wrongly estimated. Because of this, the MD distance between the sample and the controls can be overestimated or underestimated. Let's assume that the eigenvalue of the first eigenvector in figure 3.2 is biased upwards and the second eigenvalue is biased downwards. This means that the SPC model in the first row in the figure will underestimate the MD in situation (a), and overestimate the distance in situation (b). Another clear disadvantage of the eigenvalue bias is that the techniques are not applicable when the number of samples is smaller than the number of variables since the precision matrix, that is, the inverse of the covariance matrix cannot be estimated.

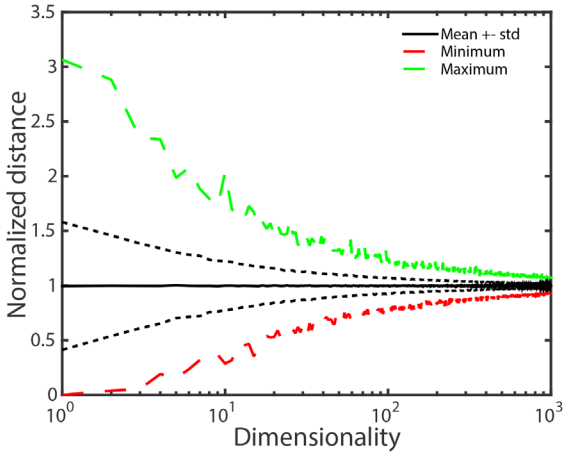

Figure 3.4 Normalized Mahalanobis distances between samples and the center of the data for different dimensionality of the data [13]. Note that for each dimensionality 1000 samples were drawn from a standard multivariate normal distribution. The distances were normalized by dividing them by the square root of the dimensionality. In this way the average distances were of comparable size over different dimensionalities.

The concentration effect

Figure 3.4 shows another effect when using the MD in high dimensional data: the distances between samples and the center of the data converge with increasing dimension [4, 13]. The same effect occurs when looking at looking at pairwise distances for any two samples in the data set. Due to this so-called concentration effect the contrast between samples diminishes as the dimension increases. This can potentially compromise the accuracy of distance-based methods. Note that the concentration effect does not necessarily hold when the samples are separated in distinct clusters. In such situations it can actually cause increased discrimination between the clusters. See Zimek et al for more details [13].

Irrelevant variables

In high dimensional data often only a few variables are related to the problem at hand while the others can be considered as irrelevant or noise variables. For example, in untargeted metabolomics data often only a few peaks (metabolites) are related to a disease while thousands are measured. As shown in figure 3.5, the performance of the MD distance is negatively influenced by the presence of irrelevant variables in many applications [13]. The effect of the relevant variables is masked by the accumulated noise of the irrelevant ones. Note that additive noise on relevant variables can cause similar problems.

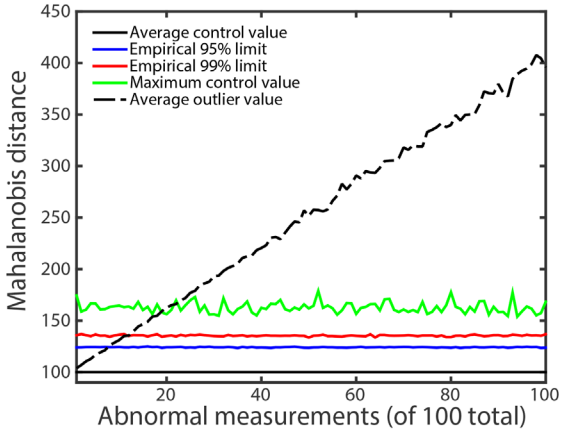

Figure 3.5 Outlier detection based on the Mahalanobis distance. This figure was generated from simulated data that followed a standard normal distribution with 100 variables [13]. The number of control and outlier samples was 10000. The outlier samples were drawn from the same distribution as the controls. Subsequently, the values of a subset of variables of the outlier samples were set to two. This means that these variables were abnormal with respect to the controls. The size of the subset was varied from 1 to 100. Note that the MD couldn't identify outliers with only a few abnormal variables.

Irrelevant variables can cause additional issues in high-dimensional modeling. First, the influence of noisy irrelevant variables often results in overfitted models. In case of LDA, it is often observed that all samples within the same class have the same score along the discriminant direction $\mathbf{a}_{LDA}$. This means that the within-group variance is zero along this direction and a perfect separation is achieved. However, such models do not generalize well to new data indicating overfitting. This is known as the data-piling problem [28].

Finally, we would like to remark that high dimensionality also brings spurious correlation; irrelevant variables may have high sample correlations to the response (e.g. a biological outcome such as disease) [3]. Such correlation may lead to false discoveries. For example in LDA, the vector $\mathbf{a}_{LDA}$ is often inspected. High absolute coefficients indicate variables that are important for classification, i.e. a disease biomarker. Irrelevant variables can receive high coefficients due to spurious correlations. Therefore, further validation is crucial to identify biomarkers that are biologically relevant.

3.2.6 Regularization

The above discussion motivates regularization of the MD to obtain better estimators for high dimensional data. As mentioned above, regularization essentially means that some structure is imposed on the model with the aim to reduce its variance at the cost of some

bias. This results in a more robust model that is less likely to overfit. We observe that in literature two common strategies are used to regularize the MD and related techniques. The first strategy essentially regularizes the covariance matrix to directly correct for the bias in the eigenvalues. The second strategy combines the MD with variable selection to directly reduce the effects of irrelevant variables and the high dimensionality of the problem. Note that the latter approach also results in a parsimonious model that is hopefully more easily interpretable.

3.3 Regularization of the covariance matrix

In high dimensional data the sample covariance matrix is not a good estimator of the population covariance matrix [14, 15]. A (naïve) strategy to obtain a more efficient estimator can be to consider an estimator with a lot of structure imposed. This way the resulting estimate contains relatively little error, but can be misspecified and therefore severely biased. For example, an independent covariance structure can be used, i.e. all covariances are zero. This implies uncorrelated variables, which is unrealistic for metabolomics data: metabolites are not independent from each other. Another often-used strategy in chemometrics to obtain more efficient covariance estimate is to restrict data analysis to the space spanned by the eigenvectors of the covariance matrix that have non-zero eigenvalues, i.e. a Moore-Penrose inverse of the covariance matrix can be used in equation 3.1. However, in such cases relevant information may be discarded. Additionally, the bias of the eigenvalues that are retained in the model is not corrected. An example of the pseudo inverse in the case of SPC is shown in row 2 of figure 3.2. In this two-dimensional model, only the first sample eigenvalue is retained, i.e. all the data is projected onto the first PC. Therefore, no difference is observed in panel b.

Many more advanced methods for regularization of the covariance matrix have been proposed in the literature [14, 15]. In general these procedures obtain an improved covariance estimator by reducing its variance at the cost of some bias. Essentially, these approaches can be divided in two categories. Many methods aim to impose special structure on the covariance matrix or its inverse such as a sparse structure, thereby reducing the variance of the resulting estimator. However, whether such structure does indeed exist cannot be verified from the data. Therefore, in some applications, a structure free approach can be preferred. Here the eigenvectors of the sample covariance matrix are retained, but the eigenvalues are corrected by pulling the highest eigenvalues downwards and the smallest ones upwards. Note that these estimators can often be interpreted as a weighted average between the sample covariance matrix and a matrix with a lot of structure (often a diagonal matrix). An example of the use of such an estimator in the case of SPC is indicated in the

third row of figure 3.2. Note that the eigenvalues are indeed adjusted as indicated by the different shape of the sphere.

In this chapter we focus on these structure free approaches. They are known as shrinkage estimators. In section 3.3.4 we provide a brief overview of estimators that do assume a special structure.

3.3.1 Steinean shrinkage

Shrinkage of eigenvalues is the oldest approach to regularization of the covariance matrix. These methods aim to correct the distorted eigenvalue structure of $\widehat{\mathbf{\Sigma}}$ and were introduced by Charles Stein (hence the name Steinean shrinkage) [15, 29]. The term shrinkage refers to the fact that these methods essentially pull (shrink) the eigenvalus towards a central value thereby correcting their bias.

Consider the spectral composition of the sample covariance matrix given by:

$$\widehat{\mathbf{\Sigma}} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^T = \sum_{i=1}^{p} \lambda_i \mathbf{e}_i^T \mathbf{e}_i \tag{3.9}$$

where $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues λ_i and $\mathbf{P}$ is an orthogonal matrix of normalized eigenvectors $\mathbf{e}_i$ (row vector). This decomposition is familiar from techniques such as PCA, where $\mathbf{e}_i$ are the principal components and $\boldsymbol{\lambda}_i$ is the explained variance. Steinean shrinkage approaches aim to regularize the covariance matrix by applying a function $\theta(.)$ to the eigenvalues such that the bias of the eigenvalues is corrected. They are of the following general form [14]:

$$\widehat{\mathbf{\Sigma}} = \mathbf{P}\theta(\mathbf{\Lambda})\mathbf{P}^T = \sum_{i=1}^{p} \theta(\lambda_i) \mathbf{e}_i^T \mathbf{e}_i \tag{3.10}$$

Note that these estimators retain the eigenvectors of the sample covariance matrix. Therefore, they are rotation invariant: rotating the variables results in the same rotation being applied to the covariance matrix estimator. Regularization with rotation equivariance is preferable when no a priori information about the structure of the covariance matrix is available. Note that although shrinkage estimators retain the sample eigenvectors this does not mean that they assume that these eigenvectors are estimated without error. It only means that it is unclear how the estimation of the eigenvectors should be improved without

imposing a specific structure on the eigenvectors or covariance matrix itself [30]. In these cases only the bias in the eigenvalues can be corrected.

A huge body of literature on shrinkage approaches, i.e. functions $\theta(.)$, is available [14, 15]. Most of these methods have been obtained by minimizing a so-called loss function. These loss functions evaluate the difference between the covariance matrix estimate and the true covariance matrix in different ways. Popular loss functions include [14, 15]:

$$\begin{aligned} L_1\left(\widehat{\mathbf{\Sigma}}, \mathbf{\Sigma}\right) &= tr\left(\widehat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1}\right) - log\left|\widehat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1}\right| - p \\ L_2\left(\widehat{\mathbf{\Sigma}}, \mathbf{\Sigma}\right) &= p^{-1}\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|_2^2 \end{aligned} \tag{3.11}$$

Loss L_1 is essentially the Kullbach-Leibler divergence between two multivariate normal densities with the same mean and covariance $\widehat{\mathbf{\Sigma}}$ and $\mathbf{\Sigma}$. Loss L_2 uses the Frobenius norm ($\|.\|_2$) to quantify the difference between the true covariance matrix and the estimator.

3.3.2 Linear shrinkage

Traditional shrinkage approaches shrink the eigenvalues towards a central value. A popular method is the linear shrinker introduced by Ledoit and Wolf [31, 32]. The following shrinkage function is applied:

$$\theta(\lambda_i)^{LW} = \delta\bar{\lambda} + (1-\delta)\lambda_i \tag{3.12}$$

where $\bar{\lambda}$ indicates the average sample eigenvalue. In figure 3.6a, it can be seen that the shrunken eigenvalues are much closer to the population eigenvalues compared to those of the sample covariance matrix. As shown in figure 3.6b, the Ledoit Wolf (LW) estimator shrinks the sample eigenvalues towards a central value (one in this case) in a linear way. The parameter δ varies between zero and one and determines the amount of shrinkage that is applied. When $\delta = 1$ the sample eigenvalues are used and when $\delta = 0$ all eigenvalues are set to the average eigenvalue. The optimal amount of shrinkage is most likely somewhere in between these extremes and can, for example, be determined by cross-validation. Theiler et al developed an extremely fast cross-validation procedure for this purpose [33]. However, such an approach is relatively computer intensive. Ledoit and Wolf showed that an analytical expression for the optimal amount of shrinkage can be obtained by minimizing loss L_2 with respect to δ [31, 32]. Schafer and Strimmer argue that in practical problems the following solution is obtained [34]:

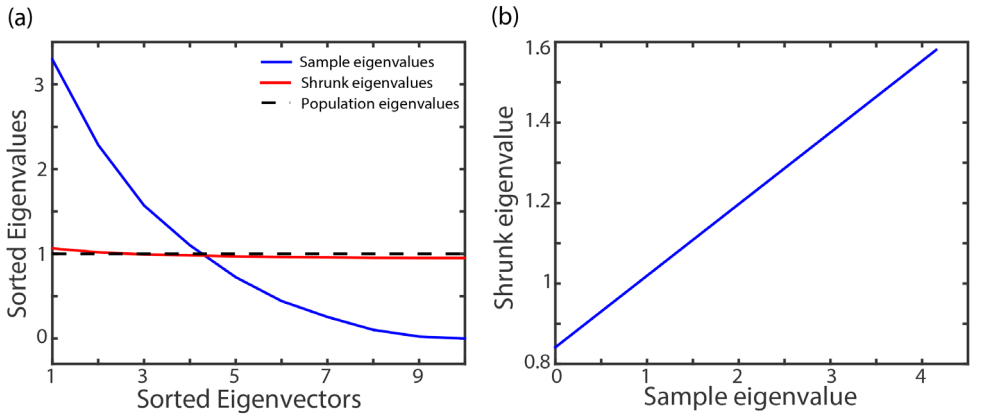

Figure 3.6 Comparison of the sample and LW covariance estimators for simulated data with 10 samples and 10 variables. **(a)** The average eigenvalues of the estimators computed over 100 simulations. The population eigenvalues are indicated by the dotted line. **(b)** The sample eigenvalues against the shrunk eigenvalues for one simulation.

$$\hat{\delta} = \frac{\sum_{i \neq j} \widehat{Var}(\widehat{\mathbf{\Sigma}}_{ij}) + \sum_i \widehat{Var}(\widehat{\mathbf{\Sigma}}_{ii})}{\sum_{i \neq j} \widehat{\mathbf{\Sigma}}_{ij}^2 + \sum_i (\widehat{\mathbf{\Sigma}}_{ii} - \bar{\lambda})^2} \tag{3.13}$$

where Σ_{ij} corresponds to the element in the ith row and jth column of the sample covariance matrix, and $\widehat{Var}(\Sigma_{ij})$ indicates the estimated variance of that element. More details regarding estimation of the variance of $\widehat{\mathbf{\Sigma}}$ are provided in appendix A of Schafer and Strimmer [34]. In practical problems with a limited number of samples, expression 3.13 can exceed one and even become negative. In order to avoid such problems the estimated shrinkage intensity $\hat{\delta}$ is truncated to lie between zero and one. Fisher et al provide a more accurate estimation of δ for the case that the data come from a multivariate normal distribution [35]. Warton et al showed that the LW shrinkage estimator can also be obtained via a penalized likelihood approach [36]. In this case the optimal shrinkage parameter can be determined by cross-validation.

Note that the LW shrinkage estimator as in expression (3.12) can also be obtained by taking a linear combination between the sample covariance matrix and a scaled identity matrix [31]:

$$\widehat{\mathbf{\Sigma}}^{LW} = \delta \frac{tr(\hat{\Sigma})}{p} \mathbf{I} + (1 - \delta) \widehat{\mathbf{\Sigma}} \tag{3.14}$$

where **I** is the identity matrix of appropriate dimension, and $tr(\widehat{\mathbf{\Sigma}})/p$ is equal to the average sample eigenvalue. Equation 3.14 gives another interpretation of the LW estimator. It is a compromise between the sample covariance matrix and the scaled identity matrix $(tr(\hat{\Sigma})/p)\mathbf{I}$. The sample covariance matrix is unbiased but estimated with high variance, while the identity matrix estimator has little variance and possibly high bias. Therefore, one can view the LW estimator as a weighted average of a variance and bias term. The weights are determined by the data and are chosen to optimize the bias-variance tradeoff. Besides the scaled identity matrix in 3.14 other, possibly more realistic, targets can be used as well. For example, Schrafer and Strimmer introduced six different targets and their associated optimal shrinkage value such as shrinkage towards towards $diag(\widehat{\mathbf{\Sigma}})$ [34].

3.3.3 Nonlinear shrinkage

Recently, Ledoit and Wolf showed that linear shrinkage works well when p/n is large or the population eigenvalues are similar in magnitude. However, when the ratio p/n is small or the population eigenvalues are dispersed (e.g. a few PCs describe most variation in the data), linear shrinkage only improves upon the sample covariance matrix slightly [30, 37, 38]. An example is shown in figure 3.7a. The poor performance of the LW estimator is due to the fact that the same amount of shrinkage is applied to every eigenvalue. For example, if the shrinkage intensity is 0.5 every eigenvalue is moved halfway towards the grand mean of all sample eigenvalues. As shown in figure 3.7a, application of a nonlinear shrinkage function to the sample eigenvalue, where each eigenvalue is shrunk by a different amount, is a great route to further improve the shrinkage estimator.

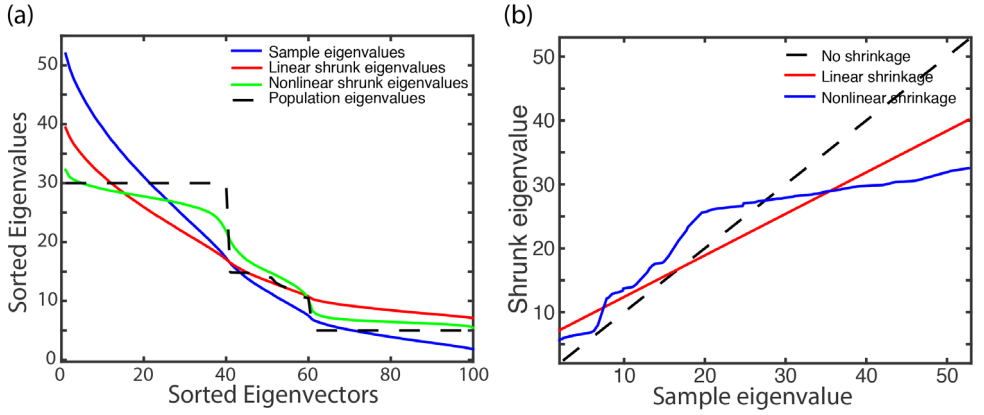

Figure 3.7 Comparison of the sample, linear shrinkage (LW) and nonlinear shrinkage (NERCOME) covariance estimators for simulated data with 400 samples and 100 variables. **(a)** The average eigenvalues of the estimators computed over 100 simulations. The population eigenvalues are indicated by the dotted line. **(b)** The sample eigenvalues against the shrunk eigenvalues for one simulation.

Recently, Ledoit and Wolf proposed such a nonlinear shrinkage estimator using random matrix theory [30, 37, 38]. We will abbreviate this method as NONLIN. The NONLIN estimator is based on a generalization of the Marcenko and Pastur equation, which describes the relationship between the sample and population eigenvalues. Paul et al provide an excellent review on the Marcenko-Pastur equation and more random matrix theory in statistics [39]. Ledoit and Wolf used the Marceno and Pastur theory to develop the so-called QuEST function (Quantized Eigenvalues Sampling Transform), which is used in the NONLIN estimator to correct the bias in each eigenvalue separately [37]. Ledoit and Wolf showed that the NONLIN method outperforms the linear LW estimator in many situations [30]. Unfortunately, their work is mathematically quite involved and difficult to digest. This hampers correct implementation of their approach by applied scientists. Additionally, the bias correction of each eigenvalue requires nonconvex optimizations making the method computationally quite intensive.

Recently, Lam proposed a non-parametric covariance estimator and showed that the eigenvalues of their estimator are asymptotically the same as those of NONLIN [40]. We will abbreviate their method as NERCOME, as in Nonparametric Eigenvalue-Regularized Covariance Matrix Estimator. The NERCOME method does not require any prior knowledge regarding the structure of the covariance matrix or its eigenvalues. The NERCOME method is based on splitting the data. The first split is used to estimate the eigenvectors of the covariance matrix. The second split is used to estimate to eigenvalues given the estimated eigenvectors. This way, the following estimator is obtained:

$$\widehat{\mathbf{\Sigma}}^{NERCOME} = \mathbf{P}_1 diag\left(\mathbf{P}_1^T \widehat{\mathbf{\Sigma}}_2 \mathbf{P}_1\right) \mathbf{P}_1^T \tag{3.15}$$

where the columns in matrix $\mathbf{P}_1$ contain the eigenvectors estimated from the first split, and $\widehat{\mathbf{\Sigma}}_2$ is the sample covariance matrix of the second split. Note that $diag\left(\mathbf{P}_1^T \widehat{\mathbf{\Sigma}}_2 \mathbf{P}_1\right)$ is an estimate of the eigenvalues based on the eigenvectors of the first split and the covariance matrix of the second split. The NERCOME method makes use of the fact that $\mathbf{P}_1$ and $\widehat{\mathbf{\Sigma}}_2$ are independent to regularize the eigenvalues. Lam argued that the split used to estimate $\widehat{\mathbf{\Sigma}}_2$ should contain max $(30, \frac{n}{10})$ samples, and that the other samples should be used to estimate $\mathbf{P}_1$. Additionally, they showed that the estimator in 3.15 can be further improved by averaging the covariance estimates over many random splits of the data. Typically, averaging over 50 splits is sufficient. These settings were used to generate figure 3.7. Lam et al show that the NERCOME estimator is optimal with respect to Frobenius loss (an expression similar to L_2 is used) when estimating the covariance matrix, and optimal with respect to Steins loss (L_1) when it is used to estimate the precision matrix. Lam et al compared the performance of NERCOME, NONLIN, and the method of principal

orthogonal complement thresholding (POET) which is briefly mentioned in section 3.3.4 [40]. The nonlinear LW and NERCOME methods were compared for simulated data where 40% of the population eigenvalues were small, 20% had intermediate values and 40% were large. In this setting, the three approaches had similar performance. The NERCOME, NONLIN, and POET approach were also compared for the case where the covariance matrix has an underlying factor structure. It was shown that NERCOME outperforms NONLIN and POET in many situations in that case. Since Ledoit and Wolf showed that NONLIN outperforms the linear LW method, it is expected that NERCOME does so as well.

Recently, two penalized likelihood approaches were proposed that essentially shrink the sample eigenvalues in a nonlinear fashion. Below, the resulting shrinkage functions are presented. Won et al proposed a condition number regularized (CNR) estimator [41]. The condition number is the ratio between the largest and smallest eigenvalue of the estimated covariance matrix. The penalty on the condition number results in a the following shrinker:

$$\theta(\lambda_i)^{CNR} = \begin{cases} \tau^*, & \lambda_i \leq \tau^* \\ \lambda_i, & \tau^* < \lambda_i < \kappa_{\max}\tau^* \\ \kappa_{\max}\tau^*, & \lambda_i \geq \kappa_{\max}\tau^* \end{cases} \tag{3.16}$$

where $\tau^* > 0$. The value for τ^* and κ_{max} are determined from the data by cross-validation. From equation 3.16 it is clear that the CNR estimator truncates the sample eigenvalues larger than $\kappa_{\max}\tau^*$ to $\kappa_{\max}\tau^*$, and those smaller than τ^* to τ^*. The CNR estimator gives most improvement compared to linear shrinkage when a few eigenvalues are much larger than the others. These effects diminish when this is not the case.

Chi et al use a nuclear norm penalty that steers the eigenvalues away from the extremes 0 and ∞ [42]. This way the so-called Covariance Estimate Regularized by Nuclear Norms (CERNN) shrinker is obtained:

$$\theta(\lambda_i) = \frac{-n + \sqrt{n^2 + 4\tau\alpha[n\lambda_i + \tau(1-\alpha)]}}{2\tau\alpha} \tag{3.17}$$

The parameter τ is determined by cross-validation where α are chosen as follows:

$$\hat{\alpha} = \left[1 + \left(\frac{1}{p} tr\left(\hat{\boldsymbol{\Sigma}}\right)\right)^2\right]^{-1} \tag{3.18}$$

The parameter τ essentially determines the amount of shrinkage that is applied while α defines a prior where the eigenvalues are shrunk towards. By means of expression 3.18 the eigenvalues are shrunk towards the grand mean of the sample eigenvalues. Compared to the CNR method, the CERNN method shrinks extreme eigenvalues in a similar manner but less drastically and shrinks intermediate eigenvalues similarly to linear estimators such as the LW estimator. Chi et al compare the linear LW, CNR and CERNN estimators with respect to loss L_1 and a quadratic loss function [42]. They show that CNR performs best when only very few population eigenvalues have a large value. When a larger fraction of population eigenvalues is large the CERNN method performs best. Ledoit and Wolf also observed that NONLIN outperforms CNR and LW in this situation. However, the NONLIN and CERNN methods have not been compared so far.

The properties of the estimators discussed above are summarized in table 3.1. Note that only a rough discrimination between the properties of the methods is provided in the table. This is due to the fact that the comparison studies in literature focus on different loss functions and only compare a few methods in a limited number of scenarios. The methods presented in this section comprise only a small set of all the shrinkage approaches that have been proposed in literature. We chose these methods to demonstrate different approaches and because we think they are interesting and can be easily implemented. Many more approaches have been developed. We would like to point the interested reader to a fascinating paper by Dohono et al who proposed nonlinear shrinkers for 26 different loss functions for spiked covariance matrices, i.e. a covariance matrix with only a few large eigenvalues [43]. Many additional methods are referenced in this paper. Additional references can also be found in [14, 15].

Table 3.1 An overview of the properties of five eigenvalue shrinkage approaches. Note that three different eigenvalue structures of the population covariance matrix are considered. In a bulk structure all eigenvalues are similar in size. In a spiked structure a few eigenvalues are larger than the others. A block structure indicates that blocks of eigenvalues of similar size can be detected, i.e. one block with low values, a block with intermediate values, and a block with high values.

		Eigenvalue structure			
Method	Shrinkage-type	Bulk	Spiked	Block	Computational complexity
LW	Linear	Good	Poor	Poor	Low
NONLIN	Nonlinear	Good	Unknown	Good	High
NERCOME	Nonlinear	Good	Unknown	Good	Medium
CNR	Nonlinear	Poor	Good	Poor	Medium
CERNN	Nonlinear	Good	Unknown	Good	Medium

3.3.4 Other regularization approaches

In this chapter we presented several shrinkage approaches for regularization of the covariance matrix. As mentioned above, many other regularization approaches have been proposed in literature. In contrast to shrinkage of eigenvalues, these approaches impose special structure on the covariance matrix. Additionally most (not all) methods are computationally more demanding. A brief overview of these other regularization methods is presented below. We focus on the main categories of methods, more details can be found in numerous references such as [14, 15].

Penalized likelihood approaches

Many regularization approaches can be interpreted as penalized likelihood estimators [14, 36, 41, 42, 44, 45]. Essentially, these methods estimate the covariance matrix via maximum likelihood, but include a penalty to the likelihood function to reduce the variance of the estimate. Many of the linear and non-linear shrinkage estimators described above can be expressed in the framework of penalized likelihood. However, penalization offers a wider range of regularization options besides shrinkage. In recent years there has been a great deal of interest in estimation of sparse covariance or precision matrices (i.e. many elements in the estimate are zero). Such a structure of the precision matrix is of interest since a zero indicates conditional independence between the variables in this case. A sparse structure is often enforced by using a so-called ℓ_1-norm penalty, i.e. the sum of the absolute values of the elements of the estimate has to be lower than a constant [44, 45]. As will be shown in section 3.4, this penalty has the property that many elements are set exactly to zero.

Banding, tapering and thresholding estimators

Banding, tapering and thresholding are another class of methods that assume that the covariance matrix is sparse [14, 15, 46, 47]. A big advantage of these methods is their low computational complexity. These methods assume that the covariance matrix is subdiagonally sparse (banding, tapering) or in its individual elements (thresholding). Sparsity is enforced by weighting the diagonals of the matrix or shrinking its individual elements to zero. Banding and thresholding estimators are not necessarily positive definite. Positive definiteness can be guaranteed for tapering estimators.

Factor models

There is a wide literature on assuming a low dimensional structure of the covariance matrix [14, 15]. This structure is typically modeled by a factor model. Recently Fan et al introduced principal orthogonal complement thresholding (POET) [48]. The method combines a factor

model estimator with thresholding estimators. More specifically they described the covariance matrix as a sum of two matrices, namely an estimate of covariance based on the first k principal components and an estimate based on the $(p-k)$ residual components. The latter estimate was regularized by means of thresholding. Note that the resulting POET estimator seems well suited for data that does not completely correspond to a factor or a sparse structure.

Regularization of the eigenvectors

There is also considerable interest in estimation of sparse eigenvectors of the covariance matrix. This essentially results in a sparse PCA model with easily interpretable loading vectors. Algorithms have been obtained by Witten et al and Shen et al, for example [49, 50]. Note that the shrinkage approaches discussed in this chapter are also of use in the context of PCA to obtain more accurate estimates of the explained variance per PC [37].

Final remarks

Many of the methods discussed above can be recast as problems involving linear or generalized linear models. An excellent overview of the use of covariance estimation via the generalized linear model (GLM) perspective is presented in the following references [14, 15]. Although difficult to digest these papers are worth looking into since the use of the GLM with its well-developed theory may offer new insights for estimation of covariance matrices.

Similarly, a Bayesian framework also offers the advantage of regularization in a well-defined statistical approach and deserves future attention. A good introduction into this topic is presented in the review by Pourahmadi et al [15]. With the advance of markov chain monte carlo (MCMC) techniques we expect that this method can also be used for data encountered in chemometrics and metabolomics.

3.4 Variable selection

The MD is hard to implement when dimensionality is high due to the difficulty of estimating the unknown covariance matrix. As mentioned in section 3.2.6, replacing the sample covariance estimator by a regularized estimator can regularize the MD. This way the bias in the eigenvalues is corrected. In the context of LDA, Bickel et al showed that assuming variable independence (i.e. a diagonal covariance matrix) could improve classification accuracy when the dimension is large [18]. Fan et al, however, showed that

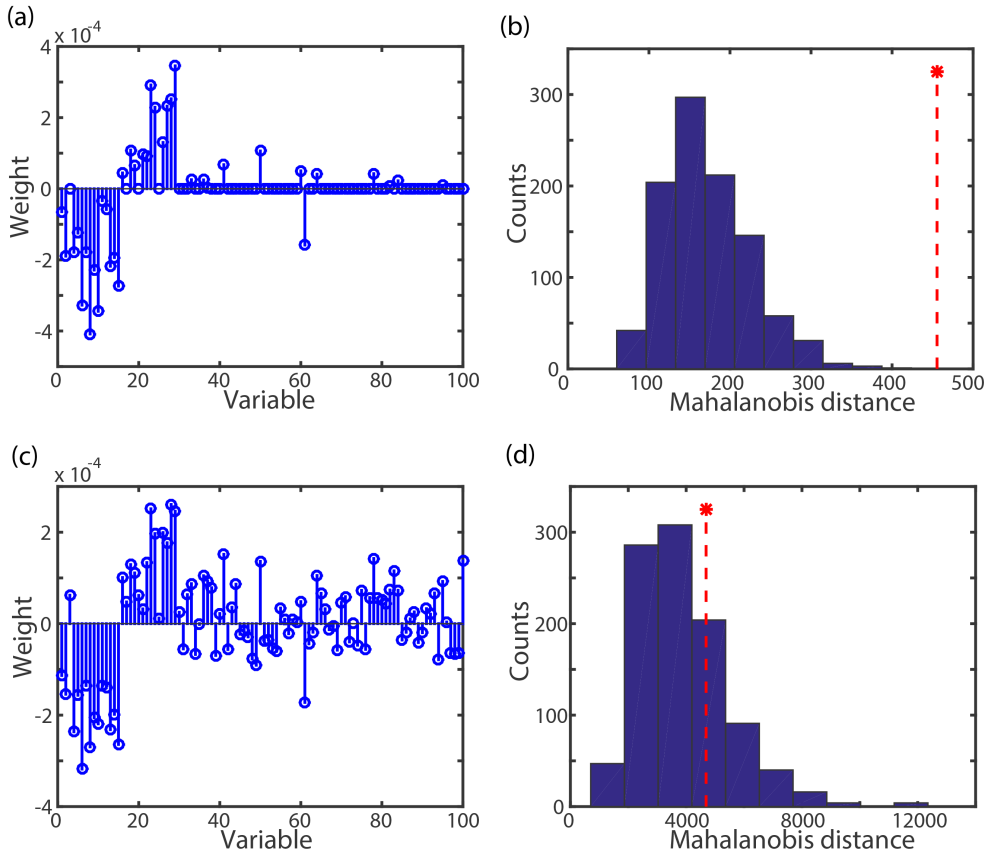

Figure 3.8 Comparison of outlier detection based on the full and sparse Mahalanobis distance for simulated data with 50 training samples and 100 variables. The model used here is introduced in chapter 6. It is a constrained model that makes vector $\mathbf{a}_{SPC}$ sparse. **(a)** The loading that was obtained for a sparse test with 30 nonzero (selected) variables. **(b)** The distribution of the distances of the control samples (blue) and the outlier or patient (red) that were found for this test. In panels **(c)** and **(d)** the same results are shown, but then for a test that used all variables.

even with the independence assumption classification using all variables could be as bad as random guessing due to noise accumulation issues [51].

Often only a few variables are related to the problem at hand, i.e. a few metabolites are related to a biological outcome. Therefore, variable selection is another powerful approach to regularization of the MD [16, 52]. Variable selection directly improves the MD's by excluding irrelevant variables from the analysis. Another advantage of this approach for metabolomics is that the resulting model is more easily interpretable. This is useful for biomarker discovery. An example in the context of SPC is shown in figure 3.8. Note that

most variable selection methods focus on selection of a subset of variables that is useful to build a good predictor, i.e. good classification accuracy in LDA [52]. This is not the same as finding all variables that are biologically related to the outcome. A subsampling approach where the results from many variable selections are combined can often be used to find most biologically relevant variables [52, 53].

Variable selection has a long history in chemometrics and many methods have been proposed. See, for example, the following excellent reviews [6, 16, 21, 52]. We present a selective overview of methods relevant to this thesis in this section. First, we briefly summarize some subset selection approaches. The rest of the section focuses on recent developments involving constrained and penalized models. This is an important class of variable selection approaches originating from the field of statistics [6, 16, 21]. Interestingly, these methods have been much less widely embraced by the metabolomics and chemometrics communities. They do, however, deserve more exploration in these fields due to their interesting (mathematical) properties [54, 55].

3.4.1 Subset selection approaches

Best subset selection is the most exhaustive approach to variable selection. Here, the model of interest is fitted to each combination of the p variables to identify the subset of relevant/important variables. Typically, criteria such Akaike's information criterion (AIC) or the Bayesian information criterion (BIC) are used to identify the best subset [6, 21]. Another approach is cross-validation. A wide range of smart search strategies such as branch-and-bound, simulated annealing, and genetic algorithms can be used to find the best subset [52]. Most search strategies are computationally expensive and not applicable in high dimensions. However, some efficient methods have been proposed that inspect only a limited number of subsets [6, 16, 21, 52]. Forward selection and backward selection are popular approaches [6, 21]. Note however, that these methods give nested models. In that sense forward and backward selection might be suboptimal compared to best subset selection [6, 21]. For example, the best predicting set with two variables does not necessarily include the best univariate predictor. Additionally, backward selection can often not be applied when $n << p$.

3.4.2 Constrained/penalized models

The variable selection methods that were discussed above fit models that contain only a subset of the variables. An alternative strategy is to fit a model on all variables, but use a technique to constrain or regularize the model parameters. In other words, use constraints that shrink the coefficients towards zero. This greatly reduces the variance of the estimator

since the model becomes less flexible due to the constraint. Additionally variables are implicitly selected as those variables having a nonzero coefficient. Constrained models have the following general form [16]:

$$L(\beta), \text{ subject to} \sum_{j=1}^{p} P(\beta_j) \leq \tau \tag{3.19}$$

where β indicate the model coefficients, $L(\beta)$ is a loss function (i.e. a least squares objective or likelihood function), $P(.)$ specifies the constraint, and τ controls the strength of the constraint. The parameter τ can be optimized in different ways, for example by cross-validation. For SPC, LDA, and MANOVA the loss function is either formulated from the maximum likelihood or the Rayleigh quotient interpretation (see section 3.2). By either minimizing or, depending on the loss function, maximizing objective 3.19 we hope to simultaneously estimate the coefficients of the variables and shrink them to zero. In other words, it is not required to fit a new model at each step as in subset selection methods. For example, the SPC example in figure 3.8 was generated penalizing the estimation of $\mathbf{a}_{SPC}$; a so-called ℓ_1-norm constraint was used (see below). Estimation of $\mathbf{a}_{SPC}$ is a minimization problem where a direction $\mathbf{a}$ is sought that minimized the variance of the controls, subject to that the distance between the "outlier" and the controls is one.

Usually constrained models are specified in Lagrangian or penalized form [56]:

$$L(\beta) + \lambda \sum_{j=1}^{p} P(\beta_j) \tag{3.20}$$

where λ is a Lagrange multiplier. Note that the constraint is now added to the objective function and acts as a penalty to obtain a bias-variance tradeoff. It can be shown that when the problem in equation 3.19 is convex (concave), the solution to minimizing (maximizing) 3.19 and 3.20 is the same [56]. Additionally, this solution is the global optimum when the so-called Kurush-Kuhn-Tucker (KKT) conditions are fulfilled [56]. Via the Lagrangian form it is possible to determine the solution to equation 3.19 analytically. This way, very fast and efficient algorithms to solve constrained models have been developed [57, 58].

Ideally, the ℓ_0-norm is used as a constraint, which is just the number of variables that are included in the model [16]. This essentially gives rise to a best subset selection method. The ℓ_0-norm constraint can also be related to model selection criteria such as AIC and BIC [16]. Unfortunately, problems with an ℓ_0-norm constraint are very difficult to solve since the

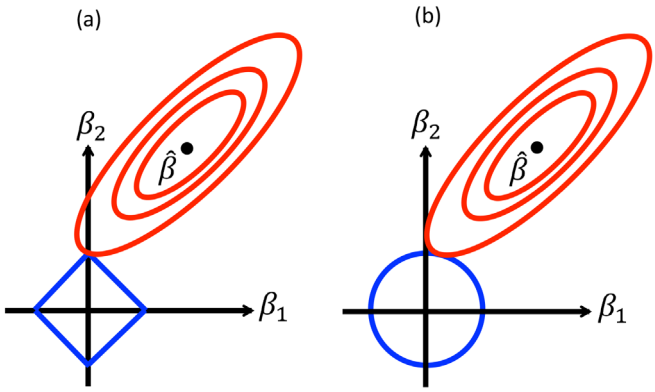

Figure 3.9 Geometric interpretation of **(a)** ℓ_1-norm constrained and **(b)** ℓ_2-norm constrained problems.

constrained problem is not convex. Therefore, a broader class of penalties is often considered, namely ℓ_q penalties where $0 \leq q \leq 2$ [6, 16]. Although all these penalties shrink the model coefficients, it can be shown that penalties with $q \leq 1$ have the sparsity property, i.e. they perform variable selection. Additionally all penalties with $q \geq 1$ are convex which often allows for an analytical solution. Because of this the ℓ_1-norm, that is, the sum of the absolute values of the coefficients, is a popular constraint since it is the convex constraint that is closest in nature to the ℓ_0-norm. In figure 3.9 it is shown why this constraint results in variable selection, while, for example, an ℓ_2-norm (the sum of squared coefficients) does not.

This figure is a geometric representation of the constrained loss in equation 3.19. The unconstrained optimal solution is marked by $\widehat{\boldsymbol{\beta}}$. The red isocontour lines represent regions of suboptimal solutions with constant loss. The blue square and circle indicate ℓ_1, and ℓ_2 constraints, respectively. Note that a solution to 3.19 is found when $\widehat{\boldsymbol{\beta}}$ falls inside the area of the constraint. In that case the unconstrained solution is obtained. When this is not the case, the solution is found where the constraint and isocontour lines intersect. With an ℓ_2-norm this will generally not occur on an axis since the constraint has a circular form with no edges. The ℓ_1-norm, however, "sticks out" at each of the axis, and so the isocontour lines will often intersect with the constraint exactly at an axis. When this occurs the coefficient of the other axis (corresponding to a variable in the data set) will be exactly zero. This induces variable selection. In this example, variable 2 is selected by the model and variable 1 receives a coefficient of zero.

As mentioned above, the strength of the constraint is determined by τ. This parameter

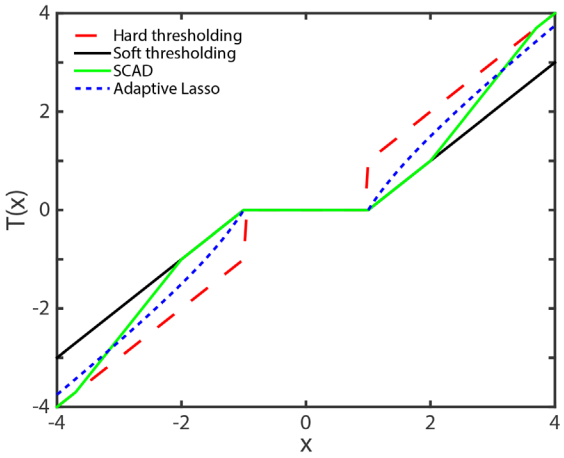

Figure 3.10 Coefficients x plotted against their constrained (penalized) values $T(x)$ for different constraints.

controls the area of the constraint, i.e. the size of the blue square and circle in figure 3.8. By making τ larger, the constraint increases in size. In the most extreme case, the solution $\widehat{\boldsymbol{\beta}}$ falls inside the region of the constraint and the unconstrained model is obtained. In contrast, when τ is small a very sparse solution is obtained. Note that constrained models also have a Bayesian interpretation by which their properties can be explained [6, 59].

Although ℓ_1-constrained models have the attractive property of variable selection, it is well known that they are slightly biased as well [6]. This is because the constraint shrinks all coefficients towards zero even if they correspond to relevant variables. To gain more insight in this phenomenom we consider the case where the variables are uncorrelated. In this case, the solution to the ℓ_1-constrained model often turns out to be a soft-thresholding (ST) of the coefficients of the unconstrained model [6, 21]. As shown in figure 3.10, ST shrinks all coefficients towards zero independent of their size. An ℓ_1-constrained model often compensates for this bias by choosing a slightly less sparse solution [6]. Because of this, extremely sparse solutions are not always obtained. Many other penalties have been proposed in literature to correct the bias of ℓ_1-constraints [6, 16]. These methods shrink large coefficients less severely than smaller ones. Two examples are shown in figure 3.10. The adaptive lasso is basically a weighted ℓ_1-norm constraint. This can be obtained by application of an ℓ_1-norm constrained model to scaled data. As shown in figure 3.10, the resulting penalty indeed penalizes large coefficients less strongly. Often an initial rough fit of the data by the model that uses the adaptive lasso is used to determine the weights. Another popular penalty is the smoothly clipped absolute deviation (SCAD) [16]. This penalty can directly be applied without first fitting a full model and also penalizes high

coefficients less strongly. However, the criterion is not convex, which makes computation much more difficult. Note that the hard thresholding rule is also included in figure 3.10. Best subset selection can be interpreted as hard thresholding (HT; variables are either in or out) [6]. HT is generally undesirable for high dimensional data since a small change in the data often results in disproportionally large changes in the model that uses HT. Therefore, the more continuous constraints such as the ℓ_1-norm are preferred

3.5 Applications

The regularization techniques that we discussed above are general and can be applied to many chemometrics methods. For example, nonlinear shrinkage can be used to obtain better estimates of the explained variance of principal components. An ℓ_1-norm penalty can be used to obtain sparse PCs [49, 50]. Examples of regularization in regression, PCA, PLS, and SVM can be found in [54, 55, 60].

Below we focus on regularization of the MD in the context of SPC, LDA and MANOVA since these are the main techniques that we mainly applied in this thesis. The aim is not to give an exhaustive list of all methods that have been proposed in literature, but rather to illustrate the regularization approaches discussed in sections 3.3 and 3.4 with a few applications.

3.5.1 Process control / Outlier detection

Process control techniques are often applied in an industrial setting where samples can be readily measured in large quantities [23, 24]. In such cases regularization of the MD is not required since $n >> p$. However, sometimes PCA is used to reduce the dimension of the data because only variation in a specific subspace has to be monitored. This can be seen as a form of regularization where the squared MD is used the monitor the samples in the space of the k selected PCs and the Euclidean distance is used to monitor the residual space [23, 24]. A similar decomposition is used by robust PCA models to detect outliers [25, 61]. Note that in robust PCA the principal components are only fitted on the most similar data points and therefore their estimator is most likely not influenced by the outliers in the data [61]. Although the (robust) PCA-based model is very powerful for process control and identification of outlying samples it can sometimes not correctly identify which variables cause the outlying behavior. This is due to the so-called smearing effect [62, 63].

The literature on variable selection in SPC is limited. Some initial work is reviewed in [26, 64, 65]. Recently, Zou et al and Capizzi et al combined the MD with variable selection [64,

65]. More specifically they aim to estimate the squared MD in equation 3.3 subject to an ℓ_1-norm penalty on $\mathbf{x}_p$ (the data is assumed to be centred to $\boldsymbol{\mu}_c$). Optimization of the strength of the penalty is a difficult task since crossvalidation cannot be used here. Additionally, poor performance was observed when criteria such as AIC or BIC were used. Therefore, a simple standardization of the test statistic was proposed to be able to determine models with different amounts of variables [64, 65]. For an industrial application they showed that the resulting model had more power compared to tests that were based on all variables in the data set. Additionally, these techniques avoid the smearing effect since no dimension reduction step is involved. Note that these methods assume that the ”outlying” sample is shifted in a few variables with respect to the mean of the controls. Therefore, direct penalization of the difference between these vectors is an intuitively sound choice. However, as shown in section 3.2.2 it is actually the weighted difference between these vectors, $\mathbf{a}_{SPC}$, that should be inspected for this purpose. Mai et al make a similar argument for the case of LDA (see below) [66]. Therefore we propose an SPC method with a penalty on $\mathbf{a}_{SPC}$ in chapter 6.

3.5.2 Linear discriminant analysis

It is well known that LDA is not applicable to high-dimensional or highly collinear data [6]. A common approach in chemometrics is to apply PCA to the data and perform classification in the reduced space. However, it can be that the components that were not included in the analysis are in fact related to the response. PLS-DA models are a popular alternative for classification [6]. However, interpretation and variable selection in such models is also not without its drawbacks [67].

Regularization of the covariance matrix has a long history in LDA. Already in 1995 Hastie proposed penalized discriminant analysis [68]. This method regularizes the covariance matrix in a similar way to the LW shrinker. The optimal amount of shrinkage was determined by cross-validation. Bickel et al considered an LDA model with variable independence by assuming that $\boldsymbol{\Sigma}_W$ is diagonal [18]. They showed that this model had better prediction accuracy in high dimensions then a model that used all variables. Fan et al, however, showed that even with such strict regularization, the method does no better then random guessing when the dimension of the data is extremely high [51]. Therefore, many sparse LDA methods have been proposed. These methods differ in the assumptions made on the within-group covariance matrix (diagonal or full structure), the number of groups that can be discriminated (2-group or multi-group), and the vector that is penalized (difference between means or direction $\mathbf{a}_{LDA}$). The properties of the different approaches are summarized in table 3.2. All methods optimize the strength of the penalty by crossvalidation.

Table 3.2 Comparison of 8 sparse LDA methods on the basis of three properties, namely the estimate/structure of $\mathbf{\Sigma}_W$ that is used, the vector that is penalized, and the maximum number of groups that can be discriminated by the method. Note that SM indicates penalization of the weighted class means $\boldsymbol{\mu}_i{\mathbf{\Sigma}_W^*}^{-1}$, where $\mathbf{\Sigma}_W^*$ is the covariance matrix with the assumed structure and $\boldsymbol{\mu}_i$ is the mean of the ith class. DD indicates penalization of the discriminant direction $\boldsymbol{a}_{FLDA}$.

Method	Covariance structure	Penalty	Number of groups	Ref
NSC	Diagonal	SM	Multi	[69]
PFLDA	Diagonal	DD	Multi	[70]
PFLDA	Shrunk	DD	Multi	[71]
RDA	Shrunk	SM	Multi	[72]
SDA	Shrunk	DD	Multi	[73]
SLDA	Sample	DD	2	[74]
LSDA	Sample	DD	2	[66]
MLSDA	Sample	DD	Multi	[75]

The nearest shrunken centroid (NSC) approach assumes that that within-group covariance matrix is diagonal [69]. In this approach, the standardized class means ($\boldsymbol{\mu}_i\mathbf{\Sigma}_{NSC}^{-1}$) are shrunk using soft thresholding. New samples are assigned to the closest shrunk centroid based on the MD using the regularized estimate of the within-group covariance matrix. Note that this approach reduces to the regularized LDA approach of Bickel et al when no shrinkage is applied [66]. Many experiments have shown that NSC is competitive for high dimensional data.

Using sparse estimates for the standardized class means as in NSC does not guarantee that the direction $\mathbf{a}_{LDA}$ is sparse. Therefore, Witten et al proposed penalized Fisher LDA (PFLDA) where the direction $\mathbf{a}_{FLDA}$ is directly penalized [70]. Again, variable independence is assumed (diagonal covariance). Note that the PFLDA problem is not convex and convergence of the method to the global optimum is not guaranteed. Recently, Gaynanova noted that because of this the constrained and penalized solution of PFLDA do not provide the same solution and the penalized form typically selects too many variables [76]. Gaynanova et al developed a routine to solve the constrained problem [76]. This way, more sparse solutions could be obtained compared to the algorithm of Witten et al.

The above-discussed methods assume that all the variables are uncorrelated. This is clearly not realistic in many situations. Mai et al also warn that independence rules could select the wrong features [66]. They provide some intuitive examples involving a few variables. Guo et al proposed regularized discriminant analysis (RDA). This method is closely related to NSC, but no variable independence assumptions are made. Again, this method penalizes the weighted class means instead of the discriminant direction, and its feature selection

properties can be considered somewhat conservative since it often includes a large number of variables [77]. Gaynanova et al proposed to use the LW estimator of covariance in the PFLDA method [71]. This method directly penalizes the discriminant directions $\mathbf{a}_{FLDA}$. As mentioned above, this approach typically selects too many variables since the problem is not convex. Additionally, the algorithm can get stuck in local optima.

Mai et al showed that two-class LDA problems can be rewritten as a regression problem [66]. This way, direction $\mathbf{a}_{LDA}$ can be found efficiently using efficient least squares routines such as least angle regression. We refer to the method as least squares discriminant analysis (LSDA). Wu et al, proposed sparse LDA (SLDA) to penalize $\mathbf{a}_{FLDA}$. They turn estimation of $\mathbf{a}_{FLDA}$ via the Rayleigh quotient into a convex minimization problem that can be penalized. Again this method is only applicable to two-class problems. An advantage of the technique is that the covariance matrix is used during the optimization instead of its inverse. Therefore, regularization is not required although the approach appears to be numerically more stable when some regularization is used. It is unclear if SLDA could benefit from regularization of the covariance matrix using some the more advanced approaches discussed in section 3.3. Mai et al showed that LSDA and SLDA are exactly the same [78]. Also sparse discriminant analysis (SDA) is the same as LSDA in case of two groups. An advantage of this method is that it can also be applied to multi-group problems. SDA transforms the LDA problem into an optimal scoring problem (least squares like problem) that can be penalized [77]. Gaynanova et al also extended the LSDA method to the multigroup case [75, 79].

Many of the sparse LDA methods discussed above were compared by Clemmensen [77]. In general, it was observed that methods that include the correlation between the variables outperform the methods that assume a diagonal covariance structure. Additionally, they noted that techniques that directly penalize $\mathbf{a}_{LDA}$ are much easier interpretable compared to techniques that penalize the weighted class means $\boldsymbol{\mu}_i\boldsymbol{\Sigma}_W^{-1}$.

3.5.3 Multivariate analysis of variance

Many methods have been proposed for application of MANOVA to high dimensional data. One of the first proposals to deal with high dimensional data was Dempsters non-exact test. This method studies the ratio of the trace of the between matrix and the within matrix. This way no inverse of the within-group scatter matrix has to be computed. Many variants of the method have been proposed [80-82].

Another traditional approach is to apply PCA to the data and restrict the MANOVA analysis to the first few components [83]. However, it can be that the groups are separated along the

components that are ignored. Therefore, this test is suboptimal. Langsrud proposed 50/50 MANOVA. Again, PCA is applied before MANOVA and k components are selected. However, an additional 50% of the remaining $n - k - 1$ components are also included as buffer components. This way it is less likely that components that are important to the response are ignored.

Srivastava proposed to use a Moore-Penrose inverse of $\mathbf{W}$ in equation 3.7 [80]. Remember from section 3.3 that this may be considered as regularization of the covariance matrix. They showed that this method was more powerful than the non-exact tests. However, the performance of this approach is very unstable since the eigenvalues of the covariance matrix cannot be reliably estimated (see section 3.2.5). Additionally, information related to the group differences might be ignored for the same reason as discussed earlier in the context of figure 3.2.

In metabolomics, ANOVA simultaneous component analysis (ASCA) is a popular alternative to MANOVA for analysis of designed experiments [84-86]. In essence this method applies univariate ANOVA tests to each variable in the dataset and uses PCA to combine the results of all univariate tests. We will show in chapter 4 that this approach is a regularized version of MANOVA where variable independence is assumed. We will use the LW shrinker as an intuitive way to bridge the gap between ASCA and MANOVA and place both methods in a general framework.

Tsai et al have also considered replacing the estimate of $\mathbf{W}$ by the LW shrinkage estimate for analysis of microarray data [87]. They showed that this resulted in a powerful method compared to popular alternatives in the field of gene set analysis. In their paper on penalized likelihood estimation of the covariance matrix Warton et al showed that a regularized estimate of $\mathbf{W}$ resulted in a more powerful test compared to when a Moore-Penrose pseudoinverse is used [36].

Not many sparse MANOVA methods have been proposed in the literature so far. Univariate ANOVAs are often used to determine variable importance in MANOVA tests. This way possible correlations in the data are clearly ignored. Recently, Shen at all truncated the data to the best m univariate predictors and applied a multivariate test afterwards [88]. Throughout the whole procedure it is assumed that the variables are independent. In another work, Shen et al proposed a sparse MANOVA test for gene set analysis. For comparing two groups they apply soft thresholding to $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\mathbf{W}^{-1}$ to obtain a sparse model [89]. The threshold value is chosen automatically based on signal/noise considerations. We note that this does not guarantee that the best sparse subset is found this

way. For example, it can be that the two variables in the optimal sparse two-dimensional subset have low coefficients in the full model. The MANOVA thresholding mentioned above will shrink these variables quickly to zero and ignore them in the rest of the analysis. In this respect, MANOVA could clearly benefit from the ideas proposed in the more advanced sparse LDA methods.

3.6 Discussion

The high dimensional nature of metabolomics data poses specific problems to data analysis. In this chapter we showed how more stable SPC, LDA, and MANOVA models can be obtained for such data by regularization. This is achieved, either by correcting the bias in the eigenvalues of the sample covariance matrix, or by variable selection. Sometimes a combination of both methods is used. These approaches have a wide application and are not limited to SPC, LDA, or MANOVA.

The methods presented in this chapter are applicable to data with moderate dimension. However, for data with truly high dimensions these approaches are computationally too demanding. It is of great interest to investigate how to obtain stable models in these difficult scenarios. For example, in metabolomics the above-mentioned regularization methods are applicable to binned NMR data. However, most methods run into computational issues when analyzing even higher dimensional data such as untargeted LC-MS spectra of complex matrix samples (untargeted metabolomics). In fields outside of metabolomics, much interest has focused on application of a (univariate) variable screening before application of multivariate techniques in truly high dimensional data [16]. In this way the variables that are most likely irrelevant are discarded and the dimension of the problem is greatly reduced. For example, a t-test can be used to rank the predictive ability of each sample for a group of controls and a group of patients. Subsequently, only the variables with the lowest p-value are retained for further analysis. This is known as sure independence screening [90]. However, such screening does not take the correlations between the variables into account. Extensions to take into account possible correlations have been proposed as well [16].

Dimension reduction is an important class of regularization approaches that is not detailed thoroughly in this chapter. In sections 3.3 and 3.5 it is briefly mentioned that PCA can be used to regularize the covariance matrix. However, many other dimension reduction methods can also be employed. Application of a model in the lower dimensional space is an effective type of regularization since fewer parameters have to be estimated. Examples of dimension reduction techniques include independent component analysis (ICA), self-organizing maps (SOM), or even Fisher LDA before application of another classifier [6].

In conclusion, high-dimensional data brings a unique set of challenges for data analysis. Many regularization approaches have been developed to obtain more reliable estimators, also in the context of SPC, LDA, and MANOVA (see sections 3.3 – 3.5). Many methods have been proposed that either regularize the sample covariance matrix, or use variable selection. Based on simulations and real data authors claim that such methods have great potential in many fields with high-dimensional data such as metabolomics. However, each author uses his own data and simulation to compare methods. Clearly, more exhaustive simulations and evaluation of the methods on multiple data sets is required to fairly compare the approaches. This way, methods can be identified that perform well in a wide variety of situations.

3.7 Regularization in this thesis

Applications of different types of regularization approaches in fields such as metabolomics and glycomics are limited. The approaches discussed in sections 3.3 and 3.4 bring new perspectives and deserve more attention from these communities than they have received so far. Chapters 4 – 6 in this thesis deal with regularization of the MD to make it applicable for analysis of metabolomics data. We evaluate three types of regularization, either from the perspective of SPC or MANOVA.

In chapter 4, we explore the value of regularization in the context of MANOVA. We show that ASCA is actually a MANOVA model with the implicit assumption that the within-group correlations between the variables are zero. This is not a realistic assumption, but, as mentioned in this review, can be useful for analysis of high-dimensional data. For data with limited dimensionality a better model can be obtained by taking the correlation structure into account. We achieve this by combining MANOVA with the LW shrinker. This way, a model is obtained that is essentially a weighted average of the ASCA and MANOVA models.

In chapter 5 we introduce the framework of process control for disease diagnosis. This is a novel and intuitive approach to disease diagnosis that is, for example, very useful for personalized health applications (e.g. diagnosis of (rare) diseases and health monitoring). In this chapter the covariance matrix is regularized by PCA. As mentioned in section 3.5, this results in a model that can identify patient samples very well, but cannot reliably identify the abnormal metabolites in all situations due to the smearing effect.

We propose an alternative SHM model in chapter 6. This is a sparse model that combines the MD with variable selection. Unlike other methods that have been proposed in SPC, we penalize the discriminant direction $\mathbf{a}_{SPC}$ directly which should result in more accurate

variable selection. Compared to the PCA-based approach an advantage is that the smearing effect is avoided. A disadvantage is that the method is computationally more expensive compared to the PCA-based approach discussed in chapter 5.

Although each type of regularization presented in chapters 4 - 6 is only presented in the context of one specific model/application they can most likely also be applied in the context of the other models since all models are related to the MD.

References

1. Donoho, D., *High-dimensional data analysis: the curse and blessings of dimensionality*, in *American Math Society on Math Challenges of the 21st century*2000: Los Angeles.
2. Fan, J.L., R., *Statistical Challenges with high Dimensionality: Feature Selection in Knowledge Discovery.* ArXiv Mathematics e-prints, 2006.
3. Fan, J., F. Han, and H. Liu, *Challenges of Big Data analysis.* National Science Review, 2014.
4. Clarke, R., et al., *The properties of high-dimensional data spaces: implications for exploring gene and protein expression data.* Nat Rev Cancer, 2008. **8**(1): p. 37-49.
5. Buydens, L., *Towards tsunami-resistant chemometrics.* The analytical scientist, 2013. **813**: p. 24-30.
6. Hastie, T.T., R.; Friedman, J., *The Elements of Statistical Learning.* 2001, New York, NY, USA: Springer New York Inc.
7. Lange, K., et al., *Next-Generation Statistical Genetics: Modeling, Penalization, and Optimization in High-Dimensional Data.* Annual Review of Statistics and Its Application, 2014. **1**(1): p. 279-300.
8. Buydens, L., *Towards tsunami-resistant chemometrics.* 2013.
9. Lindon, J.C., J.K. Nicholson, and E. Holmes, *The handbook of metabonomics and metabolomics*. 2007, Amsterdam, The Netherlands: Elsevier.
10. Mamas, M., et al., *The role of metabolites and metabolomics in clinically applicable biomarkers of disease.* Archives of Toxicology, 2011. **85**(1): p. 5-17.
11. Ellis, D.I., et al., *Metabolic fingerprinting as a diagnostic tool.* Pharmacogenomics, 2007. **8**(9): p. 1243-1266.
12. McNiven, E.M.S., J.B. German, and C.M. Slupsky, *Analytical metabolomics: nutritional opportunities for personalized health.* The Journal of Nutritional Biochemistry, 2011. **22**(11): p. 995-1002.
13. Zimek, A., E. Schubert, and H.-P. Kriegel, *A survey on unsupervised outlier detection in high-dimensional numerical data.* Stat. Anal. Data Min., 2012. **5**(5): p. 363-387.
14. Pourahmadi, M., *High-Dimensional Covariance Estimation*. Wiley Series in Probability and Statistics, ed. D.J.C. Balding, N.A.C.; Fitzmaurice, G.M.; Goldstein, H.; Johnstone, I.M.; Molenberghs, G.; Scott, D.W. 2013, Hoboken, New Jersey, US: John Wiley & Sons.
15. Pourahmadi, M., *Covariance Estimation: The GLM and Regularization Perspectives.* Statistical Science, 2011. **26**(3): p. 369-387.
16. Fan, J. and J. Lv, *A Selective Overview of Variable Selection in High Dimensional Feature Space.* Statistica Sinica, 2010. **20**(1): p. 101-148.
17. Bickel, P., et al., *Regularization in statistics.* Test, 2006. **15**(2): p. 271-344.
18. Bickel, P.J. and E. Levina, *Some Theory for Fisher's Linear Discriminant Function, 'Naive Bayes', and Some Alternatives When There Are Many More Variables than Observations.* Bernoulli, 2004. **10**(6): p. 989-1010.
19. De Maesschalck, R., D. Jouan-Rimbaud, and D.L. Massart, *The Mahalanobis distance.* Chemometrics and Intelligent Laboratory Systems, 2000. **50**(1): p. 1-18.
20. Mardia, K.V., J.T. Kent, and J.M. Bibby, *Multivariate analysis*. 1979: Academic Press.
21. James, G., et al., *An Introduction to Statistical Learning: with Applications in R.* 2014: Springer Publishing Company, Incorporated. 430.
22. Muncey, H., et al., *MetAssimulo:Simulation of Realistic NMR Metabolic Profiles.* BMC Bioinformatics, 2010. **11**(1): p. 496.
23. Bersimis, S., S. Psarakis, and J. Panaretos, *Multivariate statistical process control charts: an overview.* Quality and Reliability Engineering International, 2007. **23**(5): p. 517-543.
24. Joe Qin, S., *Statistical process monitoring: basics and beyond.* Journal of Chemometrics, 2003. **17**(8-9): p. 480-502.
25. Møller, S.F., J. von Frese, and R. Bro, *Robust methods for multivariate data analysis.* Journal of Chemometrics, 2005. **19**(10): p. 549-563.
26. Jiang, W. and K.-L. Tsui, *A theoretical framework and efficiency study of multivariate statistical process control charts.* IIE Transactions, 2008. **40**(7): p. 650-663.
27. Searle, S.R., *Linear models*. Wiley series in probability and mathematical statistics. 1971, New York, United States of America: John Wiley & Sons, Inc.
28. Shao, J., et al., *SPARSE LINEAR DISCRIMINANT ANALYSIS BY THRESHOLDING FOR HIGH DIMENSIONAL DATA*. The Annals of Statistics, 2011. **39**(2): p. 1241-1265.
29. Stein, C., *Estimation of a covariance matrix.* Rietz Lecture, 39th Annual Meeting IMS, Atlanta, GA, 1975.

30. Ledoit, O. and M. Wolf, *Nonlinear shrinkage estimation of large-dimensional covariance matrices.* 2012: p. 1024-1060.
31. Ledoit, O. and M. Wolf, *A well-conditioned estimator for large-dimensional covariance matrices.* Journal of Multivariate Analysis, 2004. **88**(2): p. 365-411.
32. Ledoit, O. and M. Wolf, *Improved estimation of the covariance matrix of stock returns with an application to portfolio selection.* Journal of Empirical Finance, 2003. **10**(5): p. 603-621.
33. Theiler, J. *The incredible shrinking covariance estimator*. 2012.
34. Schäfer, J.S., K, *A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics.* Statistical Applications in Genetics and Molecular Biology, 2005. **4**(1).
35. Fisher, T.J. and X. Sun, *Improved Stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix.* Computational Statistics & Data Analysis, 2011. **55**(5): p. 1909-1918.
36. Warton, D.I., *Penalized Normal Likelihood and Ridge Regularization of Correlation and Covariance Matrices.* Journal of the American Statistical Association, 2008. **103**(481): p. 340-349.
37. Ledoit, O. and M. Wolf, *Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions*, 2013, Working Paper Series, Department of Economics, University of Zurich.
38. Ledoit, O. and M. Wolf, *Optimal estimation of a large-dimensional covariance matrix under Stein's loss.* University of Zurich Department of Economics Working Paper, 2013(122).
39. Paul, D. and A. Aue, *Random matrix theory in statistics: a review.* Journal of Statistical Planning and Inference, 2014. **150**: p. 1-29.
40. Lam, C., *Nonparametric Eigenvalue-Regularized Precision or Covariance Matrix Estimator.*
41. Won, J.H., et al., *Condition-number-regularized covariance estimation.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2013. **75**(3): p. 427-450.
42. Chi, E.C. and K. Lange, *Stable Estimation of a Covariance Matrix Guided by Nuclear Norm Penalties.* arXiv preprint arXiv:1305.3312, 2013.
43. Donoho, D.L., M. Gavish, and I.M. Johnstone, *Optimal shrinkage of eigenvalues in the Spiked Covariance Model.* arXiv preprint arXiv:1311.0851, 2013.
44. Friedman, J.H., T.; Tibshirani, R., *Applications of the lasso and grouped lasso to the estimation of sparse graphical models.* Technical Report, 2010.
45. Friedman, J., T. Hastie, and R. Tibshirani, *Sparse inverse covariance estimation with the graphical lasso.* Biostatistics, 2008. **9**(3): p. 432-441.
46. Bickel, P.J. and E. Levina, *Covariance regularization by thresholding.* The Annals of Statistics, 2008: p. 2577-2604.
47. Bickel, P.J. and E. Levina, *Regularized estimation of large covariance matrices.* The Annals of Statistics, 2008: p. 199-227.
48. Fan, J., Y. Liao, and M. Mincheva, *Large covariance estimation by thresholding principal orthogonal complements.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2013. **75**(4): p. 603-680.
49. Witten, D.M., R. Tibshirani, and T. Hastie, *A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis.* Biostatistics, 2009: p. kxp008.
50. Shen, H. and J.Z. Huang, *Sparse principal component analysis via regularized low rank matrix approximation.* Journal of multivariate analysis, 2008. **99**(6): p. 1015-1034.
51. Fan, J. and Y. Fan, *High dimensional classification using features annealed independence rules.* Annals of statistics, 2008. **36**(6): p. 2605.
52. Guyon, I., et al., *An introduction to variable and feature selection.* J. Mach. Learn. Res., 2003. **3**: p. 1157-1182.
53. Meinshausen, N. and P. Bühlmann, *Stability selection.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2010. **72**(4): p. 417-473.
54. Rasmussen, M.A. and R. Bro, *A tutorial on the Lasso approach to sparse modeling.* Chemometrics and Intelligent Laboratory Systems, 2012. **119**(0): p. 21-31.
55. Filzmoser, P., M. Gschwandtner, and V. Todorov, *Review of sparse methods in regression and classification with application to chemometrics.* Journal of Chemometrics, 2012. **26**(3-4): p. 42-51.
56. Boyd, S.V., L., *Convex Optimization*. 2004, New York, NY, USA: Cambridge University Press.
57. Rosset, S. and J. Zhu, *Piecewise linear regularized solution paths.* The Annals of Statistics, 2007: p. 1012-1030.
58. Friedman, J., T. Hastie, and R. Tibshirani, *Regularization paths for generalized linear models via coordinate descent.* Journal of statistical software, 2010. **33**(1): p. 1.

59. Li, B. and P.K. Goel, *Regularized optimization in statistical learning: A Bayesian perspective.* Statistica Sinica, 2006. **16**(2): p. 411.
60. Chun, H. and S. Keleş, *Sparse partial least squares regression for simultaneous dimension reduction and variable selection.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2010. **72**(1): p. 3-25.
61. Hubert, M., P.J. Rousseeuw, and K. Vanden Branden, *ROBPCA: A New Approach to Robust Principal Component Analysis.* Technometrics, 2005. **47**(1): p. 64-79.
62. Alcala, C.F. and S. Joe Qin, *Analysis and generalization of fault diagnosis methods for process monitoring.* Journal of Process Control, 2011. **21**(3): p. 322-330.
63. Van den Kerkhof, P., et al., *Analysis of smearing-out in contribution plot based fault isolation for Statistical Process Control.* Chemical Engineering Science, 2013. **104**(0): p. 285-293.
64. Zou, C. and P. Qiu, *Multivariate Statistical Process Control Using LASSO.* Journal of the American Statistical Association, 2009. **104**(488): p. 1586-1596.
65. Capizzi, G. and G. Masarotto, *A Least Angle Regression Control Chart for Multidimensional Data.* Technometrics, 2011. **53**(3): p. 285-296.
66. Mai, Q., H. Zou, and M. Yuan, *A direct approach to sparse discriminant analysis in ultra-high dimensions.* Biometrika, 2012.
67. Tran, T.N., et al., *Interpretation of variable importance in Partial Least Squares with Significance Multivariate Correlation (sMC).* Chemometrics and Intelligent Laboratory Systems, 2014. **138**: p. 153-160.
68. Hastie, T., A. Buja, and R. Tibshirani, *Penalized Discriminant Analysis.* 1995: p. 73-102.
69. Tibshirani, R., et al., *Class Prediction by Nearest Shrunken Centroids, with Applications to DNA Microarrays.* Statistical Science, 2003. **18**(1): p. 104-117.
70. Witten, D.M. and R. Tibshirani, *Penalized classification using Fisher's linear discriminant.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2011. **73**(5): p. 753-772.
71. Gaynanova, I.B., J.G.; Wells, M.T., *Supervised Classification Using Sparse Fisher's LDA.* arXiv e-prints, 2013.
72. Guo, Y., T. Hastie, and R. Tibshirani, *Regularized linear discriminant analysis and its application in microarrays.* Biostatistics, 2007. **8**(1): p. 86-100.
73. Clemmensen, L., et al., *Sparse Discriminant Analysis.* Technometrics, 2011. **53**(4): p. 406-413.
74. Wu, M.C., et al., *Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection.* Bioinformatics, 2009. **25**(9): p. 1145-1151.
75. Gaynanova, I.B., J.G.; Wells, M.T., *Simultaneous sparse estimation of canonical vectors in the p>>N setting.* arXiv e-prints, 2014.
76. Gaynanova, I.B., J.G.; Wells, M.T., *Penalized versus constrained generalized eigenvalue problems.* arXiv e-prints, 2014.
77. Clemmensen, L.H., *On discriminant analysis techniques and correlation structures in high dimensions.* Technical Report, 2013.
78. Mai, Q. and H. Zou, *A note on the connection and equivalence of three sparse linear discriminant analysis methods.* Technometrics, 2013. **55**(2): p. 243-246.
79. Gaynanova, I.K., M., *Optimal variable selection in multi-group sparse discriminant analysis.* arXiv e-prints, 2014.
80. Srivastava, M.S. and Y. Fujikoshi, *Multivariate analysis of variance with fewer observations than the dimension.* Journal of Multivariate Analysis, 2006. **97**(9): p. 1927-1940.
81. Srivastava, M.S. and M. Du, *A test for the mean vector with fewer observations than the dimension.* Journal of Multivariate Analysis, 2008. **99**(3): p. 386-402.
82. Schott, J.R., *Some high-dimensional tests for a one-way MANOVA.* Journal of Multivariate Analysis, 2007. **98**(9): p. 1825-1839.
83. Langsrud, Ø., *50–50 multivariate analysis of variance for collinear responses.* Journal of the Royal Statistical Society: Series D (The Statistician), 2002. **51**(3): p. 305-317.
84. Verouden, M.P.H., et al., *Exploring the analysis of structured metabolomics data.* Chemometrics and Intelligent Laboratory Systems, 2009. **98**(1): p. 88-96.
85. Smilde, A.K., et al., *ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data.* Bioinformatics, 2005. **21**(13): p. 3043-3048.
86. Smilde, A.K., et al., *Generic framework for high-dimensional fixed-effects ANOVA.* Briefings in Bioinformatics, 2011.

87. Tsai, C.-A. and J.J. Chen, *Multivariate analysis of variance test for gene set analysis.* Bioinformatics, 2009. **25**(7): p. 897-903.
88. Shen, Y. and Z. Lin, *Tests for a Multiple-Sample Problem in High Dimensions.* Communications in Statistics - Theory and Methods, 2013. **43**(2): p. 291-305.
89. Shen, Y., Z. Lin, and J. Zhu, *Shrinkage-based regularization tests for high-dimensional data with application to gene set analysis.* Computational Statistics & Data Analysis, 2011. **55**(7): p. 2221-2233.
90. Fan, J. and J. Lv, *Sure independence screening for ultrahigh dimensional feature space.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2008. **70**(5): p. 849-911.

4

REGULARIZED MANOVA (rMANOVA) IN UNTARGETED METABOLOMICS

Many advanced metabolomics experiments currently lead to data where a large number of response variables were measured while one or several factors were changed. Often the number of response variables vastly exceeds the sample size and well-established techniques such as multivariate analysis of variance (MANOVA) cannot be used to analyse the data.

ANOVA simultaneous component analysis (ASCA) is an alternative to MANOVA for analysis of metabolomics data from an experimental design. In this paper, we show that ASCA assumes that none of the metabolites are correlated and that they all have the same variance. Because of these assumptions, ASCA may relate the wrong variables to a factor. This reduces the power of the method and hampers interpretation.

We propose an improved model that is essentially a weighted average of the ASCA and MANOVA models. The optimal weight is determined in a data-driven fashion. Compared to ASCA, this method assumes that variables can correlate, leading to a more realistic view of the data. Compared to MANOVA, the model is also applicable when the number of samples is (much) smaller than the number of variables. These advantages are demonstrated by means of simulated and real data examples. The source code of the method is available at the following github repository: https://github.com/JasperE/regularized-MANOVA.

This chapter has been adapted from: Engel, J., Blanchet, L., Bloemen, B., van den Heuvel, L. P., Engelke, U. H. F., Wevers, R. A., & Buydens, L. M. C. (2015). Regularized MANOVA (rMANOVA) in untargeted metabolomics. *Analytica chimica acta*, *899*, 1-12 (Featured article).

4.1 Introduction

Metabolomics experiments are often carried out according to an experimental design having factors such as dose, time, age, etc. Usually an untargeted or global approach is used in which as many metabolites as possible are measured. The goal of these experiments is to relate metabolic responses to the factors of interest and test their significance. For example, Salek *et al.* used an experimental design with disease and gender as factors to determine metabolites that are related to type 2 diabetes [1]. Ressom *et al.* characterized the metabolic changes pertaining to hepatocellular carcinoma in patients with liver cirrhosis; their design included three factors, namely disease, injection, and experiment [2].

Due to the complex nature of the metabolome and the analytical techniques used to measure it, datasets are extremely large and complex. For example, NMR based untargeted body fluid metabolomics datasets contain hundreds of variables (chemical shift values), which are related to the concentration of the different metabolites in the sample. Therefore, multivariate statistical analysis is indispensable to interpret the acquired data [3]. Lately, it has been recognized that standard tools used in metabolomics such as principal component analysis (PCA) often do not extract the relevant (biological) information from designed experiments [4]. These suboptimal results originate from the fact that PCA analysis is unsupervised and focuses on the captured data as a whole, i.e. it does not differentiate between the factors and interactions specified by the experimental design. Classification methods such as Orthogonal Partial Least Squares Discriminant Analysis (O-PLS-DA) are also sometimes used. However, the resulting model can be difficult to interpret in terms of the different factors and interactions since each cell in the experimental design is modeled as a different class. Note that these classification methods focus on prediction of new samples, which is not equal to testing of statistical significance of a factor or interaction. A classification method can be useful after the statistical significance of a factor has been established.

Analysis of variance (ANOVA) is a cornerstone statistical method for the analysis of data from an experimental design [5]. It is a univariate method and therefore it cannot take the relationships (covariance) between the different variables or metabolites into account. Multivariate analysis of variance (MANOVA) is the multivariate extension of ANOVA [6]. This method does take the experimental design, as well as the covariances between the metabolites, into account. MANOVA cannot be applied to data where the number of samples is (much) smaller than the number of variables because the inverse of the within-group scatter matrix cannot be computed in this case. Such high dimensional data sets are common in metabolomics. ANOVA simultaneous component analysis (ASCA) was

presented as an alternative to MANOVA for the analysis metabolomics data from an experimental design [7]. This method combines principles from ANOVA with PCA. First, ANOVA is used to separate variation in the data into different sub-matrices according to the different factors and interactions. Next, the interesting parts are analyzed by PCA in a multivariate fashion. The loadings of the PCA model can be used to determine which metabolites are related to a particular factor or interaction. ASCA has been successfully used in a number of studies such as [7, 8]. Several variants of this approach have been proposed in different fields. Examples include Scaled-to-Maximum, Aligned, and Reduced Trajectories (SMART) [9]; Principal Response Curves (PRC) [10]; and ANOVA-PCA [11]. Recently, some of these variants were described in a general framework [12].

Because PCA is a multivariate technique, it was assumed that ASCA is so as well. The first contribution of this chapter is to show that methods based on ANOVA models – such as ASCA – do not truly analyse the data in a multivariate way. From the viewpoint of MANOVA, ASCA implicitly assumes independence of the variables and equal variance for all metabolites. It is known that metabolites may correlate with one another. Therefore, we will show that the independence assumption made by ASCA is most likely incorrect, which reduces the power of the method to detect significant factors. Moreover, because of the assumption of variable independence, ASCA can mark a significant metabolite as unimportant and a less important metabolite as important. This greatly hampers the value of ASCA for biomarker identification. As a second contribution, we show that the power of ASCA to detect significant factors can be increased by a proper scaling of the data in combination with a different test statistic. As a third contribution we introduce regularized MANOVA (rMANOVA): a method that is essentially a weighted average of ASCA and MANOVA. The optimal weight is determined in a data-driven fashion. In this way the strengths of both methods are combined: rMANOVA can take the correlation between variables into account and is also applicable when the number of variables vastly exceeds the sample size. Simulated as well as real metabolomics data sets are used to compare ASCA to rMANOVA.

4.2 Theory

Consider a metabolomics experiment of p metabolites (variables) with a one-way design with k experimental conditions of sample sizes n_k. We will refer to the set of samples measured under the same condition as a group. ASCA, MANOVA, and rMANOVA can be used to analyse such data and determine whether the groups differ significantly. Below, the theory for all methods is presented. A numerical example is included to demonstrate what some of the equations represent. The data used in this example is shown in table 4.1. Note

Table 4.1 Artificial data used to explain the ASCA and MANOVA methods.

		Concentration	
Experiment	Sample label	Metabolite 1	Metabolite 2
1	-1	1.3	1.3
2	-1	0.2	1.3
3	-1	2.4	0.1
4	+1	1.3	3.0
5	+1	2.9	2.8
6	+1	2.8	3.6

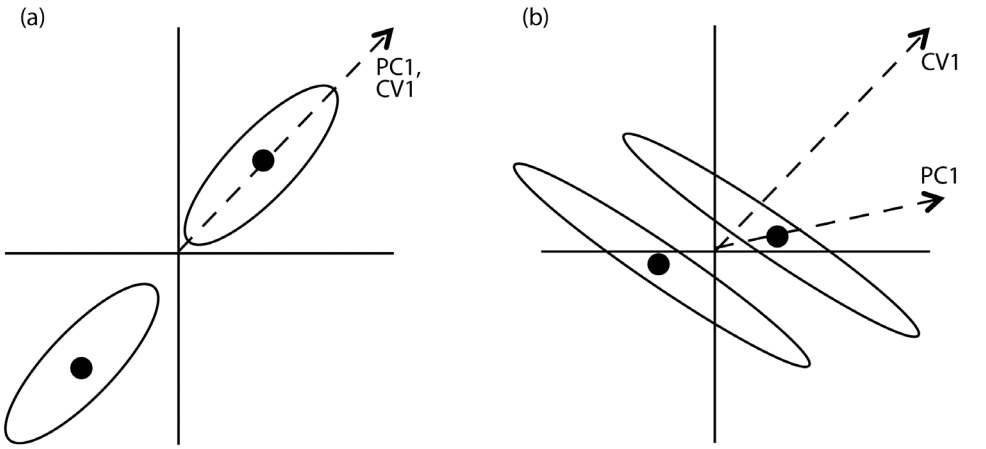

Figure 4.1 Two examples to illustrate the different approaches of ASCA and MANOVA for $k = 2$ and $p = 2$. Note that the group means are indicated by the black dots and the spread of the data around these means is represented by the ellipses. ASCA and MANOVA estimate the direction along which the groups are most separated. These are indicated by PC1 and CV1, respectively. Note that panel **(a)** corresponds to the case where the ASCA and MANOVA model give the same result. In panel **(b)** the solution found by ASCA is suboptimal and differs from that of MANOVA because the spread of the data is not taken into account.

that this example involves 2 metabolites ($p = 2$), one factor with 2 levels ($k = 2$), and three samples per level ($n_k = 3$).

4.2.1 ASCA

The aim of ASCA is (i) to determine the average metabolite concentration for each experimental condition, (ii) determine if these averages differ significantly, and (iii) apply PCA to the averages to determine which metabolites are affected by a change in the experimental conditions. An example of an ASCA model is shown in figure 4.1 for $k = 2$

and $p = 2$. Note that the model can easily be extended to more complicated designs with more effects and interactions.

Let $\mathbf{x}_{ij}$ be the vector of metabolite concentrations measured for sample j in condition i. For example, $\mathbf{x}_{12}$ corresponds to the second row or sample in table 4.1. The basis of ASCA is the linear model

$$\mathbf{x}_{ij} = \boldsymbol{\mu} + \mathbf{c}_i + \mathbf{e}_{ij} \tag{4.1}$$

Where $\boldsymbol{\mu}$ corresponds to the overall mean, $\mathbf{c}_i$is the average metabolite concentration for the i-th condition and $\mathbf{e}_{ij}$ are the residuals. Note that the data is centred by inclusion of $\boldsymbol{\mu}$ in equation 4.1. For the data shown in table 4.1, $\hat{\boldsymbol{\mu}} = [1.8\ 2.0]$, $\hat{\mathbf{c}}_1 = [-0.5 - 1.1]$, and $\hat{\mathbf{c}}_2 = [0.5\ 1.1]$. The residuals can be found by subtracting $\hat{\boldsymbol{\mu}}$ and $\hat{\mathbf{c}}_i$ from the data. Combining equation 4.1 for all samples and all factors leads to

$$\mathbf{X} = \mathbf{D}\boldsymbol{\Xi} + \mathbf{E} \tag{4.2}$$

where $\mathbf{D}$ is a design matrix that shows the labeling of the samples in the experimental design. The matrix $\boldsymbol{\Xi}$ contains the parameters of the model (4.2) and can be estimated via least squares

$$\hat{\boldsymbol{\Xi}} = (\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\mathbf{X} \tag{4.3}$$

For the numerical example in table the following matrices are obtained:

$$\mathbf{D} = \begin{bmatrix} 1 & -1 \\ 1 & -1 \\ 1 & -1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}, \hat{\boldsymbol{\Xi}} = \begin{bmatrix} 1.8 & 2.0 \\ 0.5 & 1.1 \end{bmatrix}$$

The first column in the design matrix corresponds to the overall mean ($\boldsymbol{\mu}$). The second column indicates the label of each sample for factor $\mathbf{c}$. The first and second rows in the effects matrix $\hat{\boldsymbol{\Xi}}$ correspond to $\boldsymbol{\mu}$ and $\mathbf{c}$, respectively. Note that estimates of $\boldsymbol{\mu}$ and $\mathbf{c}$ can be obtained by multiplying the appropriate columns(s) from the design matrix with the corresponding rows of $\hat{\boldsymbol{\Xi}}$. The same approach can be used for more complicated designs involving multiple factors and interactions. When the data is unbalanced, the calculations are slightly more involved. This is detailed further in appendix A. With respect to the

balanced example data, by multiplying the column in $\mathbf{D}$ corresponding to factor $\mathbf{c}$ ($\mathbf{D}_c$) with the corresponding row in $\widehat{\Xi}$ ($\widehat{\Xi}_c$), estimates for the group averages of factor $\mathbf{c}$ are obtained:

$$\mathbf{X_c} = \mathbf{D}\widehat{\Xi}_c = \begin{bmatrix} -1 \\ -1 \\ -1 \\ 1 \\ 1 \\ 1 \end{bmatrix} * [0.5 \quad 1.1] = \begin{bmatrix} -0.5 & -1.1 \\ -0.5 & -1.1 \\ -0.5 & -1.1 \\ 0.5 & 1.1 \\ 0.5 & 1.1 \\ 0.5 & 1.1 \end{bmatrix}$$

Matrix $\mathbf{X}_c$ is called a hypothesis matrix. It contains estimates of the relevant group averages for a specific factor or interaction as rows. In this case the group averages $\hat{\mathbf{c}}_i$. ASCA uses the hypothesis matrix to determine whether the corresponding factor was significant and which metabolites are related to the factor.

As an example, let's consider the problem of testing the following hypothesis:

$$H: \mathbf{X}_c = 0 \text{ vs } A: \mathbf{X}_c \neq 0 \tag{4.4}$$

The aim of this hypothesis is to determine whether factor $\mathbf{c}$ is significant (i.e. at least one of the group means $\mathbf{c}_i$ differs from the other means). The null hypothesis is that there is no effect. To test whether the null hypothesis can be rejected, ASCA uses the sums-of-squares (*SSQ*) of the hypothesis matrix as a test statistic [13]. Note that under the null hypothesis (i.e. no significant effect), the group averages are expected to be zero since the data was centred. In this case small *SSQ*-values are expected. Large differences between the means (the rows in $\mathbf{X}_c$), and therefore a large *SSQ*, are expected when an effect is significant. The distribution of the test statistic under the null hypothesis is determined by a permutation test. We will demonstrate below that alternative statistics such as Wilks' Lambda provide a more powerful test.

If the group means are found to be significantly different, PCA is applied to the hypothesis matrix (e.g. $\mathbf{X}_c$). The $k-1$ PCs are linear combinations of the mean centred variables along which the distance between the group means is the largest (see figure 4.1). In this figure ASCA actually only "sees" the 2 group averages and neglects the spread of the samples around these averages represented by the ellipses. Their (standardized) loadings can be used to determine which metabolites were affected most by the hypothesis: high absolute values of the coefficients indicate great importance.

We will show in section 4.2.3 that the ASCA model can be severely biased due to the fact that the spread of the samples around the group means is ignored. Because of this ASCA

may ascribe the wrong variables to a group difference. In some cases, a significant difference between the groups is not even detected. In this discussion it is useful to be able to show the connection between ASCA and MANOVA. An essential remark in this respect is that the PCs of $\mathbf{X}_c$ (up to a scaling factor) are actually the same as the eigenvectors of the so-called matrix of the sum of squares and products due to the hypothesis, or, simply "between" matrix. The between matrix can be easily calculated from the hypothesis matrix. For example, for factor **c** it is given by

$$\mathbf{B} = \mathbf{X}_c^T \mathbf{X}_c = \left(\mathbf{D}_c \widehat{\boldsymbol{\Xi}}_c\right)^T \left(\mathbf{D}_c \widehat{\boldsymbol{\Xi}}_c\right) \tag{4.5}$$

A numerical example of this matrix is presented in the next section on MANOVA.

As mentioned earlier, the expressions for the hypothesis matrix, and therefore **B** as well, only apply for experiments with a balanced design. For unbalanced cases, so-called sums-of-squares corrections have to be carried out to estimate the between matrix correctly. Appendix A contains more details on this point.

4.2.2 MANOVA

MANOVA is based on the same linear model (equations 4.1 and 4.2) as ASCA [6, 14]. The method assumes independency between the observations. The groups of each of the k experimental conditions are assumed to be multivariate normal with equal covariance matrices. In ANOVA the test statistic is based on a comparison of the between and within-group variance for a specific variable. In MANOVA this is not a sufficient measure since the covariance between the variables has to be taken into account. Therefore, the between-group and within-group dispersion matrices are used. The between-group dispersion matrix **B** we defined already in the equation 4.5. It represents the scatter of the group averages of a specific factor. The within-group dispersion matrix **W** is the residual sum of squares and products, or "within" matrix.

$$\mathbf{W} = \left(\mathbf{X} - \mathbf{D}\widehat{\boldsymbol{\Xi}}\right)^T \left(\mathbf{X} - \mathbf{D}\widehat{\boldsymbol{\Xi}}\right) = \widehat{\mathbf{E}}^T \widehat{\mathbf{E}} \tag{4.6}$$

The matrix **W** is a measure of the scatter of the samples within a group around their group mean as represented in figure 4.1 by the ellipses. In analogy with ANOVA, a ratio is formed between the between-group dispersion and the within-group dispersion to determine whether and effect is significant. More specifically, hypothesis tests in MANOVA are based on the matrix

$$\mathbf{J} = \mathbf{W}^{-1}\mathbf{B} \tag{4.7}$$

For the example in table 4.1, the matrices **B**, **W**, and **J** are equal to:

$$\boldsymbol{B} = \begin{bmatrix} 1.6 & 3.5 \\ 3.5 & 7.5 \end{bmatrix}, \mathbf{W} = \begin{bmatrix} 4.0 & -1.2 \\ -1.2 & 1.3 \end{bmatrix}, \mathbf{J} = \begin{bmatrix} 1.5 & 3.3 \\ 4.0 & 8.7 \end{bmatrix}$$

Note that these matrices have dimension $2\ x\ 2$ since the example only involves two metabolites.

The $k-1$ eigenvectors of **J** are linear combinations of the mean-centred variables along which the distance between the group means is the largest compared to the variation within the groups. An example is presented in figure 4.1 for $k = 2$ and $p = 2$. Large absolute coefficients of the (standardized) eigenvectors indicate metabolites that change significantly in concentration when the experimental conditions are changed.

MANOVA hypothesis tests are based on the eigenvalues (λ_j^J) of matrix **J**. These eigenvalues are a measure of the ratio of the between group dispersion and the within-group dispersion along the direction of the eigenvectors. There are four MANOVA test for significance: Wilks' Lambda, Pillai's trace, Hotellings trace and Roys greatest root [6]. These four test statistics do not always agree and their power is different under various conditions. In this work we use Wilks' Lambda (Λ) which is one of the more popular test statistics

$$\Lambda = \prod_j 1/(1+\lambda_j^J) \tag{4.8}$$

where the product runs from $j = 1$ to $j = \min(k-1, p)$. Large differences between the experimental conditions result in large eigenvalues of **J** and therefore a smaller value of Wilks lambda. The null distribution of Wilks' Lambda is often approximated by an *F*-distribution.

When the number of metabolites is larger than the number of samples (n), the inverse of matrix **W** cannot be computed and MANOVA breaks down. Essentially, the estimate of **W** (and therefore the MANOVA model) is unreliable since it suffers from high variance. Unfortunately, this is often the case in metabolomics experiments. Even when p is close to n, **W** becomes unstable in the sense that small perturbations in measurements can lead to disproportionately large fluctuations in its elements.

4.2.3 Comparing ASCA and MANOVA

The similarities and dissimilarities between the ASCA and MANOVA model are evident when considering identification of important metabolites by the methods. As mentioned above, both techniques estimate eigenvectors for this purpose, where high absolute coefficients indicate important metabolites. However, these eigenvectors are estimated from different matrices. MANOVA uses eigenvectors of $\mathbf{J} = \mathbf{W}^{-1}\mathbf{B}$, while ASCA uses eigenvectors of $\mathbf{B}$. From the viewpoint of MANOVA, ASCA assumes that the within matrix $\mathbf{W}$ is equal to the identity matrix $\mathbf{I}$. This assumption has important consequences. It implies that the concentrations of metabolites are uncorrelated to each other and that the variance of all metabolite concentrations equals a constant. This can easily be seen when comparing two groups of samples with means $\mathbf{c}_1$ and $\mathbf{c}_2$.

By definition, the direction along which these groups are most separated is given by $(\mathbf{c}_1 - \mathbf{c}_2)\mathbf{W}^{-1}$. Note that this is an eigenvector of $\mathbf{J} = \mathbf{W}^{-1}\mathbf{B}$; the MANOVA model finds an estimate for this direction. ASCA, however, assumes that $\mathbf{W} = \mathbf{I}$. Therefore, the direction found by this approach is equal to the difference between the group means $\mathbf{c}_1 - \mathbf{c}_2$. This implies that ASCA determines the metabolite importance in a univariate way ignoring the shape of the data cloud around the averages, i.e. the ellipses in figure 4.1: the importance is determined by considering each metabolite on its own. The most important metabolite is the one along which the groups are most separated. The variances around these averages are not taken into account. We will show in a simulation that the ranking can be improved by first scaling the data by the within-group variances (the diagonal elements of matrix $\mathbf{W}$). However, in this case the ranking of the metabolites is still univariate: the coefficients of the resulting eigenvector are essentially two-sample *t*-statistics.

It is known that metabolites may correlate with one another. Therefore, the metabolite independence assumption made by ASCA is most likely false and its estimate for $\mathbf{W}$ can be severely biased. This has two implications: (i) ASCA has lower power compared to MANOVA in many situations, and (ii) ASCA cannot always correctly identify the most important metabolites.

4.2.4 Regularized MANOVA

MANOVA is not applicable when $n << p$ due to high variance of the estimate of $\mathbf{W}$. ASCA is applicable in every situation, but the method is severely biased. Instead of choosing between one of these models, we suggest to combine them in the following way:

$$\mathbf{J}^* = ([1 - \delta]\mathbf{W} + \delta\mathbf{T})^{-1}\mathbf{B} \tag{4.9}$$

The matrix **T** indicates a target or prior structure of **W** and will be discussed more thoroughly below. We will refer to this model as regularized MANOVA (rMANOVA). Regularization is used in many multivariate methods to obtain efficient estimators for high dimensional data, i.e. ridge regression. The parameter $\delta \in [0, 1]$ and can be used to select a trade-off between bias and variance. The ASCA model ($\delta = 1, \mathrm{T} = \mathbf{I}$) is all bias and no variance. MANOVA ($\delta = 0$) is all variance and no bias. The optimal value for δ (δ_{opt}) is calculated in a fast and analytic way according to the Ledoit-Wolf theorem [15]. This lemma has been employed with great success in many applications such as stock market prediction, portfolio optimization, weather forecasting, economic forecasting, and genomics [16-19].

The optimum δ_{opt} depends on the target matrix **T**. The choice of the target should be guided by the expected structure of matrix **W**. In this work we will investigate two targets:

- Average variance: $\mathbf{T}_1 = \frac{1}{p} tr(\mathbf{W})\mathbf{I}$
- Unique variance: $\mathbf{T}_2 = diag(\mathbf{W})$

Target one assumes the same variance for each metabolite and is very similar to the ASCA target that has been discussed so far. The disadvantage of the target used in ASCA ($\mathbf{T} = \mathbf{I}$) is that the variance assumed by this target does not reflect the true variance in the data. The term $\frac{1}{p} tr(\mathbf{W}) = v$ in $\mathbf{T}_1$ is an estimate of the average variance in the data. The optimum δ_{opt} for this target is given by Ref. [18]:

$$\delta_{\mathrm{opt}}^{1} = \frac{\sum_{i \neq j} \widehat{Var}(w_{ij}) + \sum_i \widehat{Var}(w_{ii})}{\sum_{i \neq j} w_{ij}^2 + \sum_i (w_{ii} - v)^2} \tag{4.10}$$

where w_{ij} indicates the ijth element of **W**. It can be seen that δ_{opt} is driven by the data, expressed in $\widehat{Var}(w_{ij})$ the variance of the ijth element of **W**. The choice of the target should be guided by the expected structure of matrix **W**. The second target assumes that the concentration of each metabolite can have a different variance. The optimal value for δ_{opt} for this target is given by Ref. [18]:

$$\delta_{\mathrm{opt}}^{2} = \frac{\sum_{i \neq j} \widehat{Var}(w_{ij})}{\sum_{i \neq j} w_{ij}^2} \tag{4.11}$$

Regularization of a MANOVA model with this target has been successfully applied in the context of gene set analysis for data with a one-way design [16]. However, the properties of the method were explored far less extensively compared to the present paper. Note that complete shrinkage to target $\mathbf{T}_2$ ($\delta_{\mathrm{opt}}^{2} = 1$) gives the same model as the traditional ASCA

model applied to data that was scaled by the "variances" in $\mathbf{T}_2$, i.e. that was scaled by the diagonal elements with the within-group scatter matrix $\mathbf{W}$. As mentioned in section 4.2.3 the results of this ASCA model are closely connected to those that would be obtained by *t*-tests. Other targets than $\mathbf{T}_1$ and $\mathbf{T}_2$ can also be used, i.e. targets that incorporate prior knowledge from metabolic pathways. In this study targets $\mathbf{T}_1$ and $\mathbf{T}_2$ were selected because they demonstrate the connection between ASCA and rMANOVA. Additionally, these targets only require a few parameters to estimate. Therefore they have low variance. Because of this, the regularization in equation 4.9 offers a bias-variance trade-off.

Similar to MANOVA, eigenvalues of $\mathbf{J}^*$ are used to calculate the Wilks' Lambda statistic (Λ^*) and test for significant differences between the groups. Unfortunately, the distribution of Λ^* under the null hypothesis is unknown. A permutation test can be used to estimate this distribution from the data. In this work we use a sequential permutation test [20, 21]. Sequential tests give similar results as conventional tests, but are computationally less demanding because permutation is stopped when there seems to be little evidence against the null hypothesis. In this study the rows of data matrix $\mathbf{X}$ are shuffled in each permutation. Alternative permutation strategies for ANOVA are discussed in [22, 23]. When an effect is significant the eigenvectors of $\mathbf{J}^*$ can be used to determine the important metabolites. Often it is useful to standardize the coefficients of the eigenvectors. We will refer to the standardized eigenvectors as discriminant functions ($\mathbf{DF}$). We will use this term for eigenvectors found by ASCA, MANOVA and rMANOVA.

4.3 Methods

We used two simulation designs and real data to compare ASCA, MANOVA and rMANOVA.

4.3.1 Simulation design 1

The first simulation was carried out to highlight the implications of the metabolite independence assumption of ASCA. Data corresponding to a one-way design with two levels was simulated. The data contained five metabolites with variances [1,1,3,1,1]. The group averages were $\boldsymbol{\mu}_1 = [0,0,0,0,0]$ and $\boldsymbol{\mu}_2 = [0,2,10,0,0]$. The covariance between the metabolites was zero, except for the covariance between metabolites 1 and 2, which was set to 0.9. For each group, 500 samples were drawn from a multivariate normal distribution. The simulated concentrations of the five metabolites are shown in figure 4.2a. The samples corresponding to the first group are color coded in red, and those belonging to the second group are color-coded in blue. It is clear from the figure that the groups are not separated

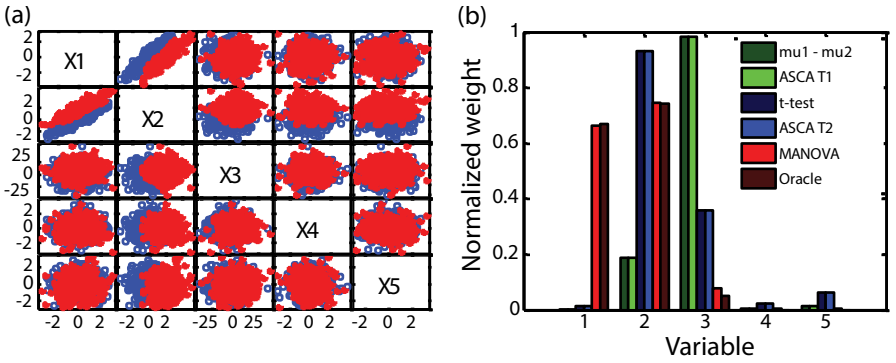

Figure 4.2 **(a)** Scatterplot of simulated metabolite concentrations for two experimental conditions or groups (red and blue) as defined in simulation 1. **(b)** The importance of the metabolites to separate these groups as determined by 5 methods. Note that the "oracle" indicates the standardized weights that are expected theoretically. The unstandardized weights are given by $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\boldsymbol{\Sigma}^{-1}$.

along metabolites 4 and 5. The groups are separated along metabolite 3 since the difference between the group averages $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ is 10 along this metabolite. However, the within-group variance is 3 and therefore the separation is still quite poor. The groups are clearly different, however, on the first two metabolites. This difference is related to the correlation between the metabolites and can only be observed when the metabolites are studied together, i.e. by a truly multivariate method. From a univariate perspective, the groups are not separated along metabolite 1 and somewhat separated along metabolite two. Based on these considerations, analysis of the data should mark metabolites 1 and 2 as most important.

4.3.2 Simulation design 2

The power of ASCA and rMANOVA tests was compared by simulation design 2. A balanced one-way design with $k = 4$ experimental conditions was considered. In each simulation the number of metabolites (p) was set to 100 and the number of samples in each experimental condition n_i was 20. Note that i indicates the ith experimental condition. The samples were drawn from multivariate normal distributions $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$.

The covariance matrix $\boldsymbol{\Sigma}$ was equal for each condition. The variance of each metabolite – the diagonal elements of the covariance matrix $(\boldsymbol{\Sigma})$ – was drawn from the uniform distribution $U(0.1,16)$. Next, these variances were combined with a specific correlation matrix $\mathbf{R}$ to form the covariance matrix $\boldsymbol{\Sigma}$. It is difficult to simulate correlation structures that closely resemble the complex structures of metabolomics data. Therefore, we studied a number of "simple" structures to be able to systematically explore the properties of the rMANOVA

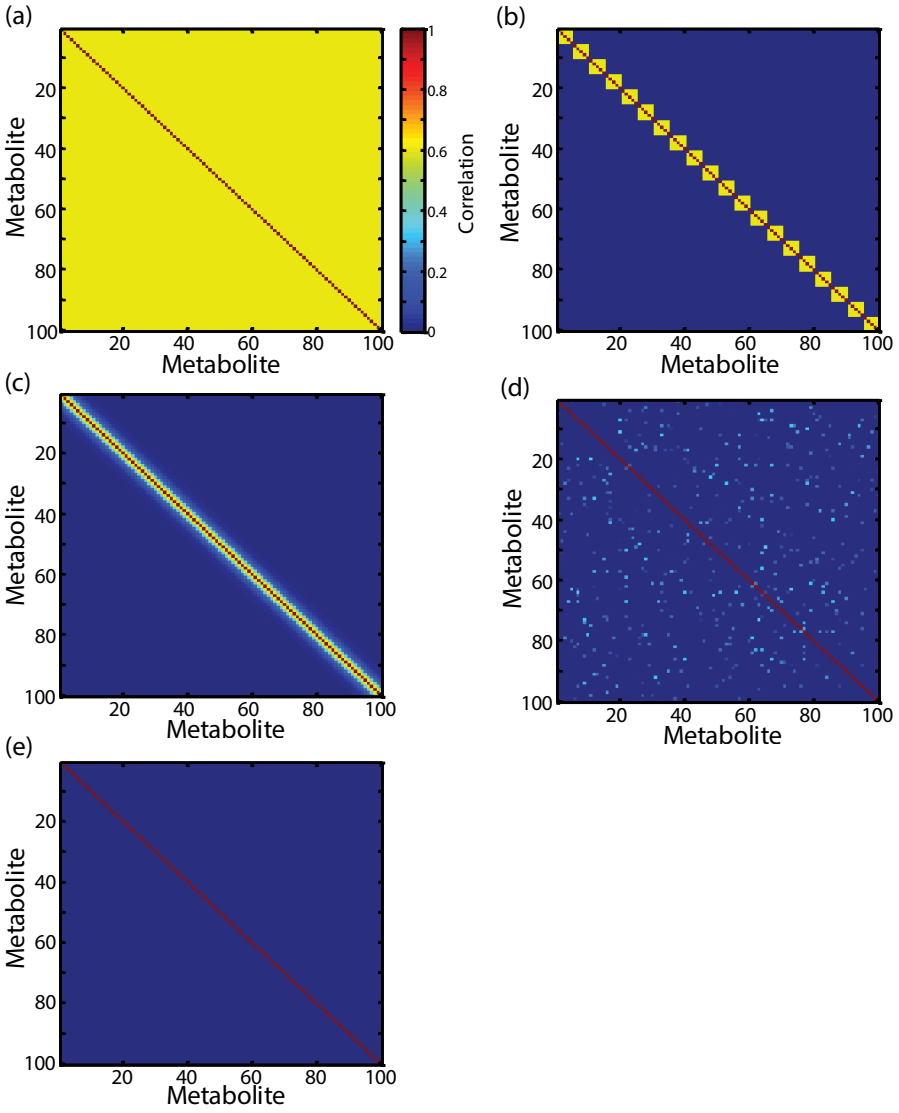

Figure 4.3 A graphical representation of **(a)** the common correlation, **(b)** the block diagonal, **(c)** the Toeplitz, and **(d)** the sparse correlation structures with $\rho = 0.6$ that were used in simulation 2. Structures **(a-c)** are equal to an independent structure when $\rho = 0$. This is shown in panel **e**.

method. Note that the simulated data contains some aspects of real metabolomics data such as correlations between variables, grouping of variables, and irrelevant noise variables. We applied rMANOVA to real NMR metabolomics data to study its performance for more complicated correlation patterns (see section 4.3.3.).

The following four correlation structures between the metabolites were considered in the simulation study:

1. **R1** Common correlation: all variables were correlated to each other with value ρ
2. **R2** Block-diagonal: variables were correlated to each other with value ρ within blocks. The different blocks were uncorrelated. Twenty blocks were defined; each block contained 5 subsequent variables. The first five variables corresponded to the first block, the next five to the second block, etc.
3. **R3** Toeplitz structure: variables close to each other were more highly correlated compared to variables that were far apart. The distance between variables with indices $p1$ and $p2$ was defined as $|p1 - p2|$. For example, the distance between the first and fourth variable in the data is equal to three. The correlation between metabolites was given by $\rho^{|p1-p2|}$.
4. **R4** Sparse structure: the word sparse here indicates that many variables were uncorrelated to each other. The correlations between the variables were drawn from a uniform distribution $U(0,1)$. This did not immediately result in a sparse matrix. Therefore, 95% of the off-diagonal elements of the matrix were set to zero. Subsequently the off-diagonal elements were divided by a constant c to ensure that the matrix was positive definite.

The following correlation values were used in structures **R1**, **R2,** and **R3**: $\rho \in \{0, 0.2, 0.4, 0.6, 0.8, 0.95\}$. Note that when the correlation was set to zero, an independent structure was obtained. A graphical representation of the different covariance structures is shown in figure 4.3.

In this simulation it was investigated how well differences between the four experimental conditions could be identified when the group means $\boldsymbol{\mu}_k$ were shifted in different ways. The following shifts were considered:

1. **S1**: The mean of each group was zero for most variables and equal to γ on five variables. For the first group, the first five variables were nonzero, for the second group the next five, etc.
2. **S2**: As **S1**, but only the first and the last group were shifted by $-\gamma$ and γ, respectively. The second and third group had zero mean on all variables. From ANOVA power studies it is known that this shift is more difficult to detect compared to shift **S1**.
3. **S3**: The groups were shifted along eigenvectors of the covariance matrix $\boldsymbol{\Sigma}$. The first group was shifted along the eigenvector with the highest eigenvalue by $\gamma\sqrt{(\lambda_1^{\Sigma})}$

units , where λ_1^{Σ} corresponds to the eigenvalue. The next group along the eigenvector with the second highest eigenvalue (λ_2^{Σ}) by $\gamma\sqrt{(\lambda_2^{\Sigma}}$) units, etc. In other words, the groups are shifted along the directions where the variation within the groups was the largest.

4. **S4**: As **S3,** but now the groups were shifted along the eigenvectors with the lowest eigenvalues. The groups were thus shifted along the directions where the variation within the groups was the smallest.

Note that for shift **S1** and **S2** the groups differ on a few metabolites. Scenarios **S3** and **S4** are extreme cases where the groups were shifted along the direction of the maximum and minimum within-group variance, respectively. The effect size γ was varied between zero and 2 in steps of 0.2. Note that $\gamma = 0$ corresponds to the null hypothesis of no significant group differences. Each simulation was repeated 1000 times. Each time, a permutation test with $m = 1000$ permutations was used to assess significance. To fairly compare ASCA and rMANOVA, both the *SSQ* and Wilks' Lambda were used to test the significance in the ASCA models. Note that MANOVA could not be applied in this simulation since it involved more variables than samples.

4.3.3 Nutritional intervention study

In this so-called Lower study, the metabolic response and loss of body fat due to four diets was studied in individuals suffering from metabolic syndrome [24]. 313 Subjects were on one of four isocaloric diets. Diets used were: (a) sustained protein normal carbohydrate; (b) sustained protein low carbohydrate; (c) supra sustained protein normal carbohydrate; and (d) supra sustained protein low carbohydrate (percentages protein : carbohydrate : fat amounted to: (a) 30 : 35 : 35; (b) 30 : 5 : 65; (c) 60 : 35 : 5 ; (d) 60 : 5 : 35). For each patient, urine samples were collected at the start of the study; after 3 months; and after 12 months. At the end of the study, each individual was flagged as responder or non-responder based on the amount of weight loss. Good responders lost more than 6% of the initial body weight while bad responders lost <6% in 3 months. NMR measurements were done on urine samples of 64 subjects, being the best and the worst responders in each of the arms of this study.

The urine samples were analysed by ^{1}H-NMR spectroscopy according to the protocol described in Wevers et al. [25]. Correlation optimized warping, histogram matching, and adaptive-intelligent binning were used for alignment, normalization and binning of the spectra, respectively [26-28]. Additionally, the bin at 3.16 ppm was removed from the data. This resonance, present in multiple samples, is associated to metformin, which is a drug that is often used in the treatment of (type II) diabetes patients. Robust PCA analyses of the data

showed four clear outlying patients because of high ethanol signal at 1.17 ppm or a high unknown peak in the glucose region around 3.53 ppm [29]. Additionally, one patient was a type I diabetic. Further analysis excluded the type I diabetic and the outlying patients. The final data set contained 59 individuals and 530 bins.

Next, ASCA and rMANOVA were used to study the contrast between the data at the start of the study and after three months. Due to the limited amount of samples, the analysis focused on the responder (**r**) and diet (**d**) effects only. Note that this corresponds to a two-way design. The following model was fit to the data:

$$\mathbf{x}_{ijk}^{t=3-t=0} = \boldsymbol{\mu} + \mathbf{r}_i + \mathbf{d}_j + \mathbf{e}_{ijk} \tag{4.12}$$

Here, index i indicates the i-th responder group, index j indicates the j-th diet, and index k the k-th sample. A permutation test with $m = 1000$ permutations was used to determine the significance of each effect. In case of significant effects the coefficients of the discriminant function were used to identify metabolites related to this effect.

4.4 Results

Below, two simulation studies and analysis of real data are used to demonstrate the properties of ASCA, MANOVA, and rMANOVA. The first simulation shows the drawback of the variable independence assumption made by ASCA and that the method offers a different – essentially univariate – view of the data. On the other hand, MANOVA offers a truly multivariate view of the data. In the second simulation the power of ASCA and rMANOVA is compared under scenarios where MANOVA cannot be applied. Finally, the value of rMANOVA compared to ASCA is also demonstrated for analysis of real metabolomics data.

4.4.1 Effect of metabolite dependence: selection of wrong variables by ASCA

In simulation 1, the concentrations of 5 metabolites and an underlying balanced one-way design with 3 levels were simulated as shown in figure 4.2a. As mentioned earlier, analysis of the data should mark metabolites 1 and 2 as most important for separation of the groups. The perfect separation can only be observed when the correlation between both metabolites is taken into account. From a univariate perspective, the groups are not separated along metabolite 1 and somewhat separated along metabolite two.

In figure 4.2b, the coefficients of the discriminant function found by MANOVA are shown. High weights indicate important metabolites. It is clear that the model marks metabolites 1 and 2 as most important. Actually, the importance values found by MANOVA are very close to the theoretical values indicated as "oracle" in the figure. The metabolite importance as given by ASCA and two univariate approaches is also shown in figure 4.2b. ASCA does not mark metabolites 1 and 2 as most important. As expected, the univariate approaches do neither. It is clear, that the results found by ASCA and the univariate approaches are exactly the same. When the average variance target ($\mathbf{T}_1$) is used, the metabolite importance is proportional to the difference of the group means. The difference between the group averages is the largest along variable 3 and therefore this metabolite is marked as most important. However, this way the variance of the data around these averages is ignored. It can be seen in figure 4.2a that the overlap between the two groups is actually quite severe. Therefore, a false interpretation of the data is obtained when the variance of each metabolite is ignored. When using target $\mathbf{T}_2$, the results are the same as those found by the two-sample *t*-test. In this case the group differences are studied relative to the variance of each metabolite. Therefore, metabolite two is marked as most important, followed by metabolite 3. The correlation between variables 1 and 2 is not observed.

This simulation shows that independence rules can select the wrong variables and essentially offer the same view of the data as univariate tests do. As a result, ASCA tests have lower power than MANOVA in many situations. Unfortunately, as mentioned above, MANOVA is not applicable when the number samples is less than the number of metabolites or variables. We propose a weighted average between the ASCA and MANOVA models. The resulting regularized MANOVA test aims to use as much information regarding between-metabolite correlations as the data allows. Compared to ASCA, this strategy deals more efficiently with situations where the number of samples is less than the number of variables.

4.4.2 The power of the tests: rMANOVA has more power in most situations

The power of the ASCA and rMANOVA tests was compared by simulation design 2. As mentioned in the methods section, the tests were compared over a wide range of correlation structures (**R1**-**R4)** and four different shifts of the group means. The patterns observed in the power simulations were similar across different correlation structures. Therefore, we mainly discuss the results for common correlation structure **R1**. All other results are provided in appendix B.

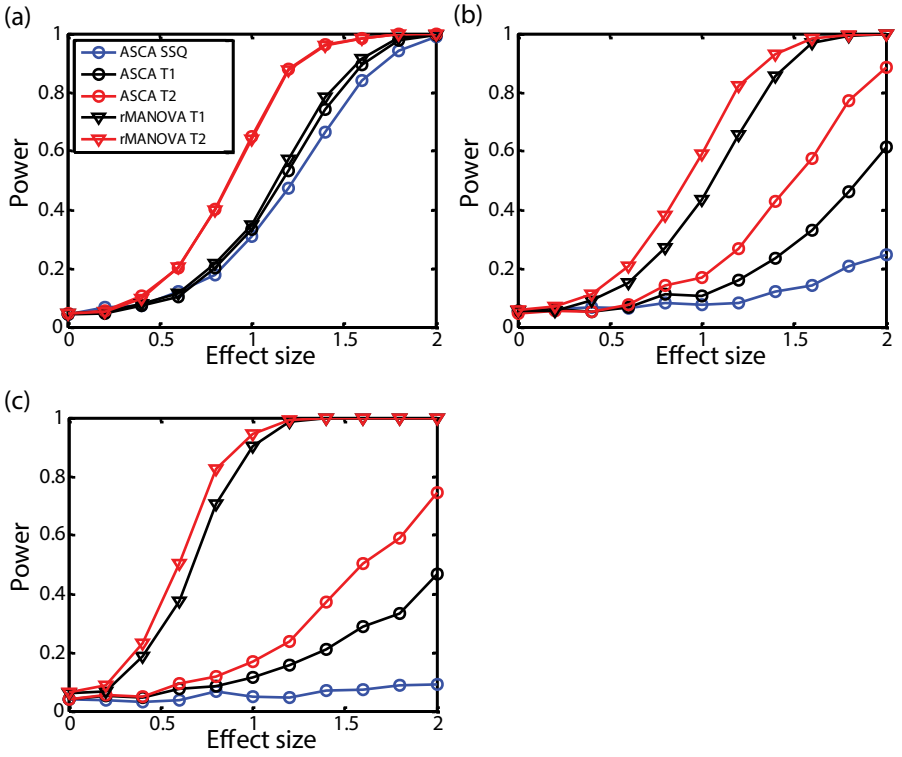

Figure 4.4 The power of the ASCA and rMANOVA tests as a function of the effect size (γ) in the case of the common correlation structure and shift direction **S1**, i.e. the groups differ on a few metabolites. Panels **a-c** correspond to $\rho = 0, 0.4$, and 0.8, respectively. Note that in panel **c** the ASCA model that uses the SSQ test has poor performance, even when the effect size is large.

Figure 4.4 illustrates the empirical powers for different effect sizes γ and different correlations ρ when the groups are shifted according to structure **S1**, i.e. the groups differ on a few metabolites. The power is defined as the percentage of simulations in which the method marked the groups as significantly different. For all methods, the type I errors are reasonably close to the chosen significance level of 5%, which can be observed when $\gamma = 0$. It can be seen that rMANOVA clearly has the highest power to detect differences between the groups ($\gamma > 0$). The largest different between the methods was observed for $\rho = 0.8$ as shown in panel c. In this case the ASCA tests had low power even when the effect was large. Only in cases of zero correlation (panel a) do the methods have similar power: here the independence assumption made by ASCA is correct. Note that for all tests the power increases with the effect size. If γ is large enough, the ASCA test can also detect a difference between the groups. In this case, prior $\mathbf{T}_2$ has the greatest power since it takes the variance

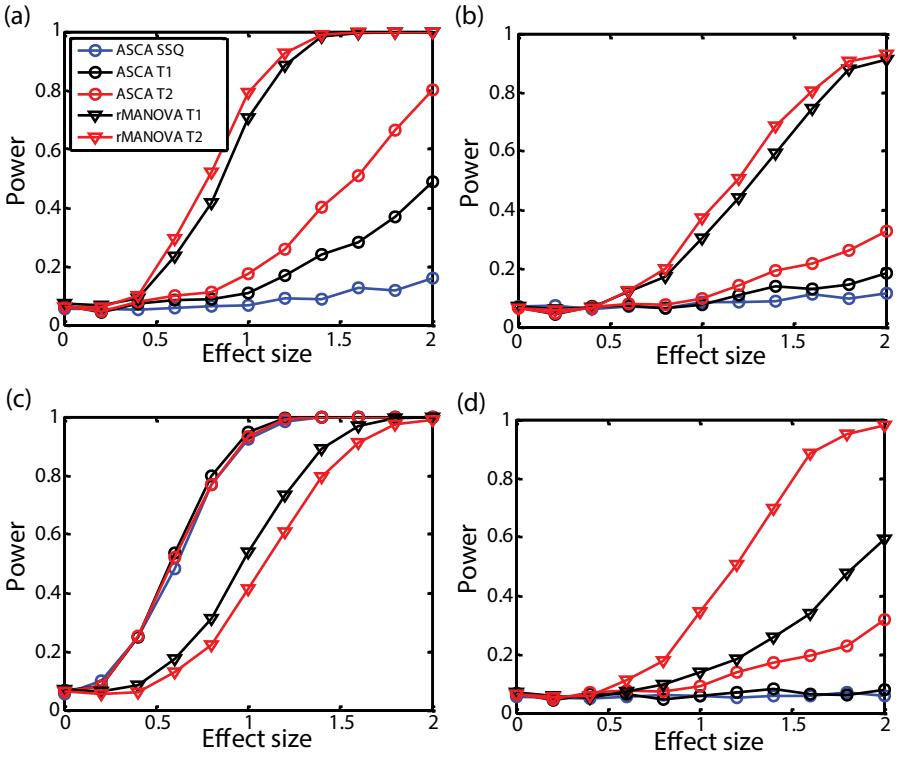

Figure 4.5 The power of the ASCA and rMANOVA tests as a function of the effect size (γ) in the case of the common correlation structure with $\rho = 0.6$. Panels **a-d** correspond to shifts **S1**-**S4**, respectively. Note that for shift **S1** and **S2** the groups differ on a few metabolites. Scenarios **S3** and **S4** are extreme cases where the groups were shifted along the direction of the maximum and minimum within-group variance, respectively.

in the concentration of all metabolites into account. Note that the power of the *SSQ* test is considerably lower compared to the other tests, which are based on Wilks' Lambda.

The power under different shifts **S1**-**S4** with $\rho = 0.6$ is shown in figure 4.5. Note that for shift **S1** and **S2** the groups differ on a few metabolites. Scenarios **S3** and **S4** are extreme cases where the groups were shifted along the direction of the maximum and minimum within-group variance, respectively. As expected, the power of the tests to detect shift **S2** was lower compared to shift **S1** since the group means were more similar. Similar to **S1**, however, rMANOVA clearly had the best power. ASCA had the greatest power when the groups are shifted in the direction of the maximum within-group variance (**S3**), as can be expected. For case **S4**, only a high power is achieved with rMANOVA using a unique variance target $\mathbf{T}_2$. In this case, target $\mathbf{T}_1$ (average variance) overestimates the variance in these directions.

Table 4.2 Standardized AUC of the size power curves for three ASCA and two rMANOVA tests for shift direction **S1**. The best performing test is indicated in boldface. Note that rMANOVA with a unique variance prior performed best in most cases.

Correlation structure	Correlation	ASCA			rMANOVA	
		SSQ	Average variance ($\mathbf{T}_1$)	Unique variance ($\mathbf{T}_2$)	Average variance ($\mathbf{T}_1$)	Unique variance ($\mathbf{T}_2$)
R1 Common correlation (Figure 4.3a)	0.00	0.42	0.45	**0.58**	0.46	0.57
	0.20	0.19	0.26	0.39	0.41	**0.44**
	0.40	0.11	0.19	0.30	0.50	**0.56**
	0.60	0.08	0.17	0.28	0.59	**0.62**
	0.80	0.06	0.16	0.26	0.68	**0.71**
	0.95	0.07	0.19	0.28	0.80	**0.83**
R2 Block Diagonal (Figure 4.3b)	0.00	0.40	0.46	**0.57**	0.45	**0.57**
	0.20	0.41	0.41	**0.55**	0.40	0.53
	0.40	0.38	0.35	**0.48**	0.36	0.46
	0.60	0.32	0.32	**0.45**	0.30	0.41
	0.80	0.26	0.28	0.41	0.31	**0.42**
	0.95	0.21	0.25	0.35	0.34	**0.48**
R3 Toeplitz (Figure 4.3c)	0.00	0.39	0.44	**0.57**	0.46	**0.57**
	0.20	0.38	0.44	0.56	0.46	**0.57**
	0.40	0.38	0.41	**0.52**	0.42	**0.52**
	0.60	0.31	0.34	0.49	0.41	**0.50**
	0.80	0.23	0.25	0.38	0.48	**0.57**
	0.95	0.09	0.17	0.27	0.69	**0.75**
R4 Sparse (Figure 4.3d)	-	0.39	0.41	**0.53**	0.43	**0.53**

In table 4.2, the AUC of the size power curves of shift **S1** is shown for all methods and all correlation structures. Similar tables for shifts **S2-4** are available in supplementary material B. In all tables, the best performing method (the highest AUC) is highlighted in boldface. Again, unique variance ($\mathbf{T}_2$) seems to outperform the *SSQ* and $\mathbf{T}_1$ tests which assume equal variance for all metabolites. Additionally, rMANOVA has the highest power for shift **S1**, **S2** and **S4**. For the sparse covariance structure (**R4**) shrinkage did not greatly improve the power. This lack of difference is due to properties of our simulation: the sparse covariance structure was very close to diagonal to satisfy the semi-positive definite constraint (see figure 4.3).

Table 4.3 The significance of factors responder and diet given by the ASCA and rMANOVA tests. The parameter δ indicates the amount of shrinkage towards the target. Note that $\mathbf{T}_1$ and $\mathbf{T}_2$ indicate an average variance and a unique variance prior, respectively.

Model	*p-value* responder	*p*-value diet	δ
ASCA SSQ	0.57	0.38	1
ASCA $\mathbf{T}_1$	0.57	0.67	1
ASCA $\mathbf{T}_2$	0.08	0.04	1
rMANOVA $\mathbf{T}_1$	0.04	0.02	0.27
rMANOVA $\mathbf{T}_2$	0.01	0.00	0.48

4.4.3 Real data

Metabolic syndrome is an important risk factor for developing diabetes or cardiovascular disease amongst others. A possible therapy is to influence the metabolism by changing food intake patterns. In this study, the metabolic response and loss of body fat due to four diets was studied in individuals suffering from metabolic syndrome. The main aim was to assess differences in metabolic response between responders and non-responders to the diet, where a responder was defined as an individual who lost more than the average amount of weight during the study.

In table 4.3, the results of the different models are shown. It can be seen that rMANOVA uses some correlation information in the model since $\delta < 1$. The ASCA tests cannot use this information. This is reflected in the p-values observed for the methods. The regularized methods (rMANOVA) mark both factors as significant, while ASCA does not. ASCA in combination with $\mathbf{T}_2$ does give p-values below a significance limit of 10%. However, when a limit of 5% is employed only the diet factor is marked as significant. In figure 4.6, the density of the samples projected onto the discriminant function for the response factor is shown. It is clear that rMANOVA separates the two groups much clearer compared to ASCA.

Since the main aim of the study was to assess the metabolic differences between responder and non-responder the absolute coefficients of the responder discriminant function were inspected to identify important metabolites. Unfortunately, the function was rather noisy and important metabolites could not easily be identified. As a first tentative approach, the 25 largest absolute coefficients of the discriminant function were used to identify important metabolites. The important metabolites included beta-hydroxybutyric acid (1.22/1.24 ppm), acetone (2.22 ppm), and acetoacetic acid (2.30 ppm), collectively referred to as the ketone bodies. These ketone bodies are related to the breakdown of fatty acids for energy supply. This process is known to be triggered during caloric restriction and during starvation.

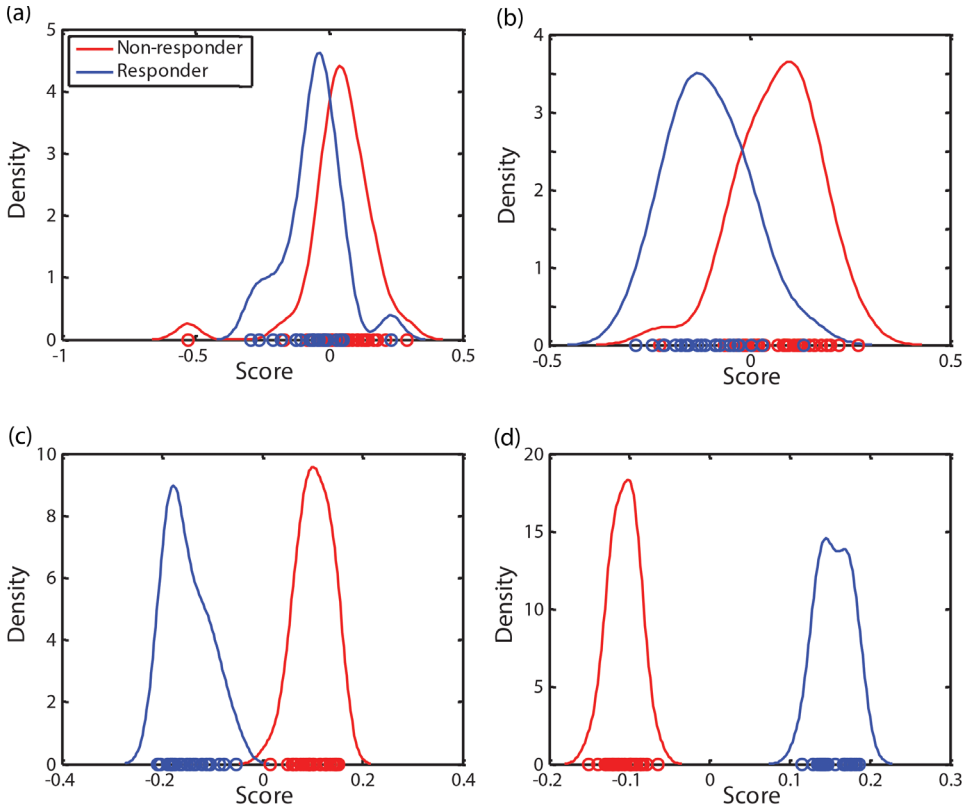

Figure 4.6 Density plots of the data projected onto the discriminant functions found by ASCA with targets **(a)** average variance ($\mathbf{T}_1$) and **(b)** unique variance ($\mathbf{T}_2$) and rMANOVA with targets **(c)** $\mathbf{T}_1$ and **(d)** $\mathbf{T}_2$.

Responders are consuming their fat reserve more than non-responders leading to an increased amount of ketone bodies. The obvious next step in the analysis is to determine the important metabolites more carefully, possibly leading to new biological knowledge. This might involve a variable selection procedure such as sure independence screening, or direct sparse estimation of the discriminant functions. Such analyses, although interesting in their own right, are beyond the scope of this article.

4.5 Discussion

All test statistics that were proposed in this study are based on Wilks'Lambda and had higher power compared to the current ASCA *SSQ* statistic. Additionally, rMANOVA had higher power compared to ASCA. In general, the power of the tests seems to increase in the order *SSQ* < ASCA $\mathbf{T}_1$ < ASCA $\mathbf{T}_2$ < rMANOVA $\mathbf{T}_1$ < rMANOVA $\mathbf{T}_2$. The lower performance

of the first three tests comes from the fact that the structure underlying the metabolite concentrations (i.e. a metabolic pathway) is ignored. However, the tests based on an independence rule had greater power when the groups were shifted in the direction of the maximum within-group variance. In practice, however, a metabolomics experiment is carried out according to an experimental design because the effect of interest is expected to be subtle. Therefore the groups are most likely not shifted in the direction of maximum variance. Additionally, such a case can most likely also be detected by regular PCA analysis of the data. Cole et al. studied how the power of MANOVA can both increase and decrease as a function of within-group correlations among the dependent variables [30]. They revealed that (a) power increases as correlations between dependent variables with effect sizes in the same direction move from near 1 to -1; (b) power increases as the correlation between variables with very different effect sizes becomes more positive or negative; and (c) power increases as correlations between dependent variables with small effect sizes shift from positive to negative. In light of these findings and our simulation results, we expect rMANOVA to be a very competitive alternative to ASCA in most situations.

In practical applications, the assumption of symmetric and equal covariance made by the methods may not hold. For example, distributions of metabolite concentrations can often be better approximated by a log-normal distribution. This can be addressed using a log-transformed version of the data. The metabolite intercorrelations in a group of diseased subjects can be very different from those in a group of healthy controls. In MANOVA literature, it is well known that Wilks' lambda is not the most robust statistics against such violations from the model assumptions [6]. It would be of interest to explore the relative merits of ASCA and rMANOVA; especially in situations where the assumptions made by the model are violated. Other approaches such as non-parametric MANOVA, Dempster's non-exact test, Srivastava's test, and the Chen-Qin test could also be of interest in this respect [31-33]. Note that these approaches do not automatically generate discriminant functions and their results are most likely less easy to interpret compared to the approach suggested in the present paper.

Often the assumption that errors are identically and independently distributed is violated as well; when analysing longitudinal data, for example. It would be interesting to use a mixed-effect model in such cases. However, it is not directly evident how the relationship between time and all the metabolites should be specified. Other possibilities include the study of contrasts (as in this paper), summary statistics, or using a repeated measurements MANOVA strategy [34].

4.6 Conclusion

In metabolomics research, it has become increasingly popular to analyse data with an underlying experimental design. In this work, we have discussed this analysis problem in the context of ASCA and MANOVA. We propose the regularized MANOVA model, which is a weighted average of the ASCA and MANOVA models. This way, correlations can be included and the model can be applied in situations when MANOVA cannot, for example the analysis of high dimensional metabolomics data. Numerical examples have demonstrated that the rMANOVA method has higher power compared to ASCA under a wide range of scenarios and offers a more realistic view of the data. Here, rMANOVA was introduced for analysis of metabolomics data, but the method can be applied for analysis of any high-dimensional data set with an underlying experimental design.

This work also brings additional insight into the ASCA method. We have shown that ASCA assumes that the metabolite concentrations are independent, and therefore the method offers a similar view of the data as univariate techniques do. Additionally, it was shown that the *SSQ*-statistic used to test significance in ASCA has low power. The power of ASCA can be improved by using the Wilks'Lambda test and by a suitable scaling of the data. However, the improved ASCA test is still suboptimal in many cases.

The rMANOVA source code is available from the first author upon request, and also at the following github repository: https://github.com/JasperE/regularized-MANOVA.

Acknowledgements

Part of the work in this paper was performed in the framework of the project "Protein-rich food against obesity" program 'Pieken in de Delta Oost-Nederland' and partly subsidized by the Ministry of Economic Affairs, Agriculture and Innovation (EL&I), and the Dutch provinces of Gelderland and Overijssel. The authors thank Dr Aart van Amerongen, PI of this project, for all his work in this context. The authors are grateful to Prof. Margriet Westerterp (Maastricht) and Prof. Bruce Wolffenbuttel (Groningen) for allowing us to use samples collected during the "Lower-study" project. The authors thank Angelina Goudswaard for performing the NMR measurements.

References

1. Salek, R.M., et al., *A metabolomic comparison of urinary changes in type 2 diabetes in mouse, rat, and human.* Physiological genomics, 2007. **29**(2): p. 99-108.
2. Ressom, H.W., et al., *Utilization of metabolomics to identify serum biomarkers for hepatocellular carcinoma in patients with liver cirrhosis.* Analytica Chimica Acta, 2012. **743**(0): p. 90-100.
3. Trygg, J., E. Holmes, and T. Lundstedt, *Chemometrics in Metabonomics.* Journal of Proteome Research, 2006. **6**(2): p. 469-479.
4. Verouden, M.P.H., et al., *Exploring the analysis of structured metabolomics data.* Chemometrics and Intelligent Laboratory Systems, 2009. **98**(1): p. 88-96.
5. Ståhle, L. and S. Wold, *Analysis of variance (ANOVA).* Chemometrics and Intelligent Laboratory Systems, 1989. **6**(4): p. 259-272.
6. Ståhle, L. and S. Wold, *Multivariate analysis of variance (MANOVA).* Chemometrics and Intelligent Laboratory Systems, 1990. **9**(2): p. 127-141.
7. Smilde, A.K., et al., *ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data.* Bioinformatics, 2005. **21**(13): p. 3043-3048.
8. Rago, D., et al., *A LC–MS metabolomics approach to investigate the effect of raw apple intake in the rat plasma metabolome.* Metabolomics, 2013. **9**(6): p. 1202-1215.
9. Keun, H.C., et al., *Geometric Trajectory Analysis of Metabolic Responses To Toxicity Can Define Treatment Specific Profiles.* Chemical Research in Toxicology, 2004. **17**(5): p. 579-587.
10. Van den Brink, P.J. and C.J.F. ter Braak, *Principal response curves: Analysis of time-dependent multivariate responses of biological community to stress.* Environmental Toxicology and Chemistry, 1999. **18**(2): p. 138-148.
11. de Haan, J.R., et al., *Interpretation of ANOVA models for microarray data using PCA.* Bioinformatics, 2007. **23**(2): p. 184-190.
12. Smilde, A.K., et al., *Generic framework for high-dimensional fixed-effects ANOVA.* Briefings in Bioinformatics, 2011. 524 - 535.
13. Vis, D., et al., *Statistical validation of megavariate effects in ASCA.* BMC Bioinformatics, 2007. **8**(1): p. 322.
14. Searle, S.R., *Linear models*. Wiley series in probability and mathematical statistics. 1971, New York, United States of America: John Wiley & Sons, Inc.
15. Ledoit, O. and M. Wolf, *A well-conditioned estimator for large-dimensional covariance matrices.* Journal of Multivariate Analysis, 2004. **88**(2): p. 365-411.
16. Tsai, C.-A. and J.J. Chen, *Multivariate analysis of variance test for gene set analysis.* Bioinformatics, 2009. **25**(7): p. 897-903.
17. Ledoit, O. and M. Wolf, *Improved estimation of the covariance matrix of stock returns with an application to portfolio selection.* Journal of Empirical Finance, 2003. **10**(5): p. 603-621.
18. Schäfer, J.S., K, *A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics.* Statistical Applications in Genetics and Molecular Biology, 2005. **4**(1).
19. Kriegeskorte, N., R. Goebel, and P. Bandettini, *Information-based functional brain mapping.* Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(10): p. 3863-3868.
20. Besag, J. and P. Clifford, *Sequential Monte Carlo p-Values.* Biometrika, 1991. **78**(2): p. 301-304.
21. Silva, I., R. Assunção, and M. Costa, *Power of the Sequential Monte Carlo Test.* Sequential Analysis, 2009. **28**(2): p. 163-174.
22. Anderson, M.J., *Permutation tests for univariate or multivariate analysis of variance and regression.* Canadian Journal of Fisheries and Aquatic Sciences, 2001. **58**(3): p. 626-639.
23. Anderson, M. and C.J.F. ter Braak, *Permutation tests for multi-factorial analysis of variance.* Journal of Statistical Computation and Simulation, 2003. **73**(2): p. 85-113.
24. Soenen, S., *Efficacy of macronutrients on targeting obesity and beyond*, 2010, Universiteit van Maastricht.
25. Wevers, R.A., et al., *1H-NMR Spectroscopy of Body Fluids: Inborn Errors of Purine and Pyrimidine Metabolism.* Clinical Chemistry, 1999. **45**(4): p. 539-548.
26. Tomasi, G., F. van den Berg, and C. Andersson, *Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data.* Journal of Chemometrics, 2004. **18**(5): p. 231-241.

27. De Meyer, T., et al., *NMR-Based Characterization of Metabolic Alterations in Hypertension Using an Adaptive, Intelligent Binning Algorithm.* Analytical Chemistry, 2008. **80**(10): p. 3783-3790.
28. Torgrip, R.J.O., et al., *A note on normalization of biofluid 1D 1H-NMR data.* Metabolomics, 2008. **4**(2): p. 114-121.
29. Hubert, M. and P.J. Rousseeuw, *ROBPCA: A New Approach to Robust Principal Component Analysis.* Technometrics, 2005. **47**(1): p. 64-79.
30. Cole, D.A.M., Scott E., Arvery, R., Salas, E., *How the Power of MANOVA Can Both Increase and Decrease as a Function of the Intercorrelations Among the Dependent Variables.* Psychological Bulletin, 1994. **115**(3): p. 465 - 474.
31. Srivastava, M.S., *On testing the equality of mean vectors in high dimension.* ACTA et Commentationes Universitatis Tartuensis De Mathematica, 2013. **17**(1): p. 31.
32. Zerzucha, P. and B. Walczak, *Concept of (dis)similarity in data analysis.* TRAC Trends in Analytical Chemistry, 2012. **38**(0): p. 116-128.
33. Anderson, M.J., *A new method for non-parametric multivariate analysis of variance.* Austral Ecology, 2001. **26**(1): p. 32-46.
34. Huberty, C.J.O., Stephen, *Applied MANOVA and Discriminant Analysis*. Wiley series in probability and statistics. 2006, Hoboken, New Jersey: John Wiley & Sons, Inc. .

Appendices

A. Sums-of-squares corrections for unbalanced data

Consider data with an underlying design with two main effects. In this case the following model is suitable:

$$x_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$$

However, there is no reason why factor α should be fitted before factor β or the other way around. The following model is also suitable:

$$x_{ijk} = \mu + \beta_j + \alpha_i + e_{ijk}$$

Actually, when analysing balanced data the sequence in which the main effects enter the model doesn't matter. Their contributions to the fit will be the same. Unfortunately, this is not the case when analysing unbalanced data. In this case the upper two models cannot be used for the same purpose. In the first model, the contribution of factor α is estimated before all other effects. Therefore, the estimated effect of α, say X_α, can be interpreted as the contribution of factor α while ignoring main effect β. In the second model, X'_α can be interpreted as the contribution of factor α given factor β ($X_{(\alpha|\beta)}$). Clearly, careful thought should be paid to which model is analysed for unbalanced data.

Studying effects given the other main effects is also referred to as type-II sums of squares correction in ANOVA. In this example, such a correction can be achieved by comparing the residual sums of squares of a full model that contained all effects ($W_{\alpha,\beta}$) to that of a reduced model that did not contain the factor of interest (W_β). The corrected sums of squares are given by $B_{(\alpha|\beta)} = W_{\alpha,\beta} - W_\beta$. Studying effects given the other main effects *and* interactions is known as type-III sums of squares.

B. The AUC of the size power curves of shifts S2-4.

Table B.1. Standardized AUC of the size power curves for three ASCA and two rMANOVA tests for shift **S2**. The best performing test is indicated in boldface. Note that rMANOVA with a unique variance prior performed best in most cases.

Correlation structure	Correlation	ASCA			rMANOVA	
		SSQ	Average variance ($\mathbf{T_1}$)	Unique variance ($\mathbf{T_2}$)	Average variance ($\mathbf{T_1}$)	Unique variance ($\mathbf{T_2}$)
R1 Common correlation (Figure 4.3a)	0.00	0.21	0.23	**0.40**	0.26	0.39
	0.20	0.11	0.12	0.19	0.21	**0.22**
	0.40	0.06	0.10	0.14	0.27	**0.34**
	0.60	0.08	0.10	0.14	0.39	**0.42**
	0.80	0.06	0.09	0.12	0.50	**0.55**
	0.95	0.07	0010	0.14	0.70	**0.73**
R2 Block Diagonal (Figure 4.3b)	0.00	0.21	0.25	**0.37**	0.24	**0.37**
	0.20	0.18	0.21	**0.36**	0.21	0.34
	0.40	0.16	0.17	**0.28**	0.19	0.27
	0.60	0.16	0.16	**0.25**	0.15	0.22
	0.80	0.12	0.15	0.23	0.17	**0.26**
	0.95	0.16	0.14	0.18	0.17	**0.28**
R3 Toeplitz (Figure 4.3c)	0.00	0.24	0.24	**0.38**	0.26	**0.38**
	0.20	0.25	0.25	**0.38**	0.24	0.36
	0.40	0.22	0.23	**0.32**	0.22	**0.32**
	0.60	0.17	0.17	0.28	0.21	**0.29**
	0.80	0.14	0.12	0.19	0.23	**0.33**
	0.95	0.07	0.10	0.13	0.48	**0.57**
R4 Sparse (Figure 4.3d)	-	0.20	0.20	**0.33**	0.26	**0.33**

Table B.2. Standardized AUC of the size power curves for three ASCA and two rMANOVA tests for shift **S3**. The best performing test is indicated in boldface. Note that ASCA performed best in many cases.

Correlation structure	Correlation	ASCA			rMANOVA	
		SSQ	Average variance ($\mathbf{T}_1$)	Unique variance ($\mathbf{T}_2$)	Average variance ($\mathbf{T}_1$)	Unique variance ($\mathbf{T}_2$)
R1 Common correlation (Figure 4.3a)	0.00	0.65	0.67	0.59	**0.66**	0.59
	0.20	**0.72**	0.71	0.70	0.56	0.47
	0.40	0.70	**0.72**	0.70	0.54	0.50
	0.60	0.70	**0.71**	**0.71**	0.53	0.47
	0.80	**0.69**	**0.69**	0.68	0.48	0.43
	0.95	**0.69**	**0.69**	0.68	0.45	0.36
R2 Block Diagonal (Figure 4.3b)	0.00	0.67	**0.69**	0.59	0.66	0.59
	0.20	**0.72**	0.71	0.67	0.69	0.66
	0.40	0.72	**0.73**	0.69	0.69	0.67
	0.60	0.73	**0.75**	0.71	0.68	0.67
	0.80	**0.75**	**0.75**	0.72	0.69	0.69
	0.95	0.75	**0.76**	0.72	0.68	0.70
R3 Toeplitz (Figure 4.3c)	0.00	**0.67**	**0.67**	0.60	**0.67**	0.60
	0.20	0.69	**0.70**	0.64	0.68	0.63
	0.40	**0.73**	**0.73**	0.68	0.69	0.67
	0.60	**0.76**	0.75	0.74	0.70	0.69
	0.80	0.77	**0.78**	0.77	0.69	0.70
	0.95	0.65	0.67	0.59	**0.66**	0.59
R4 Sparse (Figure 4.3d)	-	**0.72**	0.71	0.70	0.56	0.47

Table B.3. Standardized AUC of the size power curves for three ASCA and two rMANOVA tests for shift **S4**. The best performing test is indicated in boldface. Note that rMANOVA with a unique variance prior performance best.

Correlation structure	Correlation	ASCA			rMANOVA	
		SSQ	Average variance (T_1)	Unique variance (T_2)	Average variance (T_1)	Unique variance (T_2)
R1 Common correlation (Figure 4.3a)	0.00	0.10	0.13	**0.59**	0.16	**0.59**
	0.20	0.07	0.07	0.36	0.16	**0.41**
	0.40	0.05	0.06	0.19	0.18	**0.47**
	0.60	0.06	0.06	0.13	0.20	**0.43**
	0.80	0.05	0.05	0.07	0.16	**0.39**
	0.95	0.05	0.05	0.05	0.10	**0.3**
R2 Block Diagonal (Figure 4.3b)	0.00	0.10	0.15	**0.59**	0.15	**0.59**
	0.20	0.13	0.10	0.55	0.15	**0.56**
	0.40	0.08	0.07	0.41	0.15	**0.49**
	0.60	0.06	0.07	0.27	0.11	**0.42**
	0.80	0.05	0.07	0.12	0.11	**0.39**
	0.95	0.05	0.06	0.05	0.06	**0.16**
R3 Toeplitz (Figure 4.3c)	0.00	0.12	0.13	**0.59**	0.16	**0.59**
	0.20	0.13	0.12	**0.56**	0.15	**0.56**
	0.40	0.09	0.10	0.44	0.13	**0.48**
	0.60	0.08	0.07	0.26	0.12	**0.40**
	0.80	0.06	0.05	0.09	0.09	**0.33**
	0.95	0.05	0.06	0.06	0.07	**0.26**
R4 Sparse (Figure 4.3d)	-	0.06	0.07	0.31	0.08	**0.32**

5

TOWARDS THE DISEASE BIOMARKER IN AN INDIVIDUAL PATIENT USING STATISTICAL HEALTH MONITORING

In metabolomics, identification of complex diseases is often based on application of (multivariate) statistical techniques to the data. Commonly, each disease requires its own specific diagnostic model, separating healthy and diseased individuals, which is not very practical in a diagnostic setting. Additionally, for orphan diseases such models cannot be constructed due to a lack of available data. An alternative approach adapted from industrial process control is proposed in this chapter: statistical health monitoring (SHM).

In SHM the metabolic profile of an individual is compared to that of healthy people in a multivariate manner. Abnormal metabolite concentrations, or abnormal patterns of concentrations, are indicated by the method. Subsequently, this biomarker can be used for diagnosis. A tremendous advantage here is that only data of healthy people is required to construct the model. The method is applicable in current – population based - clinical practice as well as in personalized health applications.

In this study, SHM was successfully applied for diagnosis of several orphan diseases as well as detection of metabotypic abnormalities related to diet and drug intake.

This chapter has been adapted from: Engel, J., Blanchet, L., Engelke, U. F., Wevers, R. A., & Buydens, L. M. C. (2014). Towards the Disease Biomarker in an Individual Patient Using Statistical Health Monitoring. *PloS one*, *9*(4).

5.1 Introduction

Metabolomics is becoming increasingly important in a whole range of healthcare-related fields such as disease prevention, diagnosis and intervention, and studies of the impact of diet and nutrition on various forms of illness [1-3]. In such studies, the metabolic phenotype or metabotype of individuals is studied. The metabotype is a characteristic metabolite profile that depends on the interplay between genes and environmental factors such as diet, lifestyle, gut microbial composition, and – in patients – medication. This profile varies greatly between individuals and populations. Therefore, metabotyping has applications in population-based and personalized medicine [1, 4]. For example, various (subtle) abnormalities in the metabotype have been related to cancer states, diabetes, cardiovascular diseases, neurological diseases and inborn errors of metabolism (IEM) [1, 3, 5, 6].

Commonly, the metabotype of an individual is measured from easily accessible biofluids such as urine or serum, or more seldom from tissue [2]. Typically, untargeted metabolomics techniques such as ^{1}H nuclear magnetic resonance (NMR) spectroscopy or mass-spectrometric methods are used for this purpose [2, 7, 8]. These techniques can measure a wide range of metabolites simultaneously and generate a multivariate profile of metabolites present in the sample. Due to the complex nature of the metabolome in biofluids multivariate data analysis is often required to interpret the acquired data and detect metabolic abnormalities. Most studies deal with classification problems such as disease diagnosis (healthy versus a specific disease) [3, 9]. During data analysis, such problems are commonly tackled as a *two-class* or a *one-class* problem.

Two-class classification strategies aim to model the metabolic differences *between* groups of healthy and diseased individuals. These differences are grouped in a metabolic pattern or biomarker representing the abnormalities related to the disease. Typically used methods for two-class classification are orthogonal projection to latent structure (OPLS) and linear discriminant analysis (LDA) [9]. More challenging problems are generally tackled using non-linear approaches such as SVM, K-PLS or Random Forests [10, 11]. In contrast, one-class classification methods focus on the similarities that are encountered *within* the diseased group. This results in a characterization of the expected metabotypes for a specific disease (e.g. an average metabotype and expected metabolic variation). Here, the most commonly used technique is SIMCA [9]. For both classification strategies, a diagnosis is made by matching the metabotype of a patient against the result of the model, being this biomarker or expected metabotype.

Both strategies focus on groups of patients with one specific disease. This might be impractical in a clinical setting for three reasons. First, it is not realistic to construct a statistically reliable model for rare or orphan diseases. Such diseases are defined in the United States as any disease that affects fewer than 200000 individuals, and in the European Community as any disease that affects fewer than 5 in 10000 individuals [12]. Some rare diseases have less than a dozen known cases. In other words, the number of potential patients to base the model on is too low. Secondly, even if orphan diseases are ignored, each disease requires its own specific model. Thirdly, unknown metabolic perturbations, for instance caused by unknown diseases, may not be detected or falsely interpreted.

Interestingly, similar problems are encountered when monitoring industrial processes. Analogous to disease diagnosis one wants to know whether or not the process is in-control (healthy); if not, a known or unknown rare event (a disease) has occurred that may affect product quality. So-called statistical process control (SPC) techniques have been developed to detect all of these events as early as possible [13]. Due to the success of SPC, we propose here to adapt these strategies and apply the method on metabolome profiles of body fluids with the aim of diagnosing the disease of a patient. This provides a new tool for diagnostic support: statistical health monitoring (SHM).

In SHM, the so-called normal operating conditions (NOC) of healthy people are defined. NOC is a term that is often used in SPCA. In this case it basically means that a one-class classifier is used to model the expected metabotypes of healthy individuals. The NOC should therefore represent the average metabotype of a population and the inherent (normal) variation present in this population e.g. due to difference in life style. Next, the metabotype of a patient is compared to NOC. Deviations from NOC such as abnormal metabolite concentrations or abnormal patterns of concentrations are indicated by the method. Subsequently, this information – a (disease) biomarker for this individual patient – can be used for diagnosis. The fact that only data of healthy people is required to construct the SHM model is a tremendous advantage of this approach. Because of this, SHM is not disease specific and can be used for diagnosis of rare diseases.

As a case study we applied SHM for diagnosis of a family of orphan diseases, namely inborn errors of metabolism. IEM comprise a substantial group of rare genetic diseases that can be diagnosed by NMR spectroscopy in combination with visual inspection of the data [5]. Because of the complex structure of the spectra this can be quite a challenge. Moreover, such an approach is extremely time-consuming and quite subjective. The proposed SHM approach may make the diagnosis of IEM easier and objective. Additionally, it will be

shown that, depending on how the NOC are defined; SHM can also detect metabolic abnormalities related to diet and medication.

The next section will outline the concept of SHM and mathematic background. In the remaining sections the properties of SHM are discussed based on application of SHM to the case study example involving IEM.

5.2 Theory

In SHM the metabotype of an individual is compared to that of healthy people in a multivariate manner. Abnormal metabolite concentrations, or abnormal patterns of concentrations, are indicated by the method. This is achieved in two steps. In this first step – detection of abnormal metabotypes – the metabotype of an individual is matched against NOC and marked as normal or possibly abnormal. The abnormal metabolites are identified in a second step.

5.2.1 Detection of abnormal metabotypes

The first step in SHM is to select samples that represent the NOC of healthy humans well. From now on we will refer to these samples as normal or NOC samples. The choice of NOC samples should reflect the goal of the SHM analysis. For example, if the goal is purely to detect abnormalities related to disease, the NOC set can include healthy individuals who recently took medication. However, if one also wants to detect abnormal metabolites related to drugs, these individuals should not be included. Additionally, the demographics of the NOC samples and the expected patients should be as similar as possible. For example, if a patient has a completely different lifestyle compared to the NOC samples, many metabolites may falsely be marked as abnormal. However, if the demographics are too loosely specified, the limit of detection of the SHM model will be negatively affected. We will further elaborate on this important aspect in the discussion section.

The NOC samples are stored in data table ($\mathbf{X}_h$). Each row in $\mathbf{X}_h$ contains the metabotype information from one healthy individual. Each column corresponds to a measured feature, e.g. a chemical shift value in an NMR spectrum. The data is centred to zero mean before starting the statistical analysis. Often it is also useful to scale the data – e.g. to unit variance – to ensure that each feature has equal chance to influence the model.

Principal component analysis (PCA) is used to describe the NOC data [13]:

$$\mathbf{X}_h = \mathbf{T}_h \mathbf{P}_h^T + \mathbf{E}_h \tag{5.1}$$

Here, $\mathbf{T}_h \mathbf{P}_h^T$ is the part of the model that describes the structural metabolic variation between the NOC samples, while matrix $\mathbf{E}_h$ only contains residuals or non-structural variation. $\mathbf{T}_h$ describes the systematic metabolic differences or variation between the NOC samples. The columns in $\mathbf{P}_h$, or factors, are the actual model. The factors are descriptors that indicate in which measured features the systematic differences occur. A property of the factors is that they are ordered by importance: the first "explains" most variation, followed by the second, etc. At some point the remaining factors only describe noise. These factors are not included in the model.

To determine whether someone is possibly diseased, the metabotype information from this individual ($\mathbf{x}_{new}$) is evaluated using the constructed model:

$$\mathbf{x}_{new} = \mathbf{t}_{new} \mathbf{P}_h^T + \mathbf{e}_{new} \tag{5.2}$$

Note that $\mathbf{t}_{new} \mathbf{P}_h^T$ describes which part of the individual's metabotype is in accordance with the metabotypes that are expected for NOC samples[2]. If an individual is similar to the normal samples, this prediction should capture his/her complete metabotype. In this case, the error $\mathbf{e}_{new}$ should be small and fall in the range of the error of the NOC samples. Therefore, abnormal metabotypes can be detected by inspection of the size of $\mathbf{e}_{new}$. In industrial process control, the so-called Q-statistic is used for this purpose [13]:

$$Q = \|\mathbf{e}_{new}\|^2 \tag{5.3}$$

A sample with a high Q-value corresponds to a metabotype that either contains abnormal metabolite(s) or abnormal metabolite concentrations that break the normal between-metabolites correlation pattern. The metabotype of an individual is marked as abnormal if the value for Q exceeds the significance limit given by Q_α [14]:

$$Q > Q_\alpha \left[\frac{z_\alpha \sqrt{2\phi_2 h_0^2}}{\phi_1} + \frac{\phi_2 h_0 (h_0 - 1)}{\phi_1^2} \right]^{\frac{1}{h_0}} \tag{5.4}$$

where the parameters of the approximation are defined as:

[2] (Additional comment) Here, it is assumed that a disease will break the normal metabolite correlations, for example, by causing a block in a metabolic pathway. In the possibly unlikely scenario that correlations are not broken, but e.g. only the average concentration of certain metabolites are affected, patients may have abnormal values in $\mathbf{t}_{new} \mathbf{P}_h^T$ and not in $\mathbf{e}_{new}$ and Hotellings T^2-statistic should also be used for detection (see appendix A).

$$h_0 = 1 - \frac{2\phi_1\phi_3}{3\phi_2^2} \tag{5.5}$$

$$\phi_i = tr\left(\widetilde{\mathbf{\Lambda}}^i\right), \quad i = 1,2,3 \tag{5.6}$$

$\widetilde{\mathbf{\Lambda}}$ is the covariance matrix of $\mathbf{E}_h$, and z_α is the standardized normal variable with $(1-\alpha)$ confidence limit, having the same sign as h_0 [14].

To summarize, for a measured metabotype ($\mathbf{x}_{new}$) the PCA model ($\mathbf{t}_{new}\mathbf{P}_h^T$) is used to predict what this metabotype would look like if the individual was an NOC sample. The larger the difference between predicted and measured metabotype ($\mathbf{e}_{new}$) the more likely the metabotype is to be abnormal. The size of this difference is expressed via the Q-statistic. An example of SHM when monitoring 2 metabolites is presented in appendix A.

5.2.2 Identification of abnormal metabolites

The second step in SHM is to detect the abnormal metabolites that caused the deviation from NOC. A clinical practitioner can use this information for example for disease diagnosis, possibly via a database search.

Since the Q-statistic should detect all deviations from NOC, the contribution of measured features to this statistic should be investigated. For this purpose, the value for Q is decomposed into per feature contributions. Here, we used partial decomposition [15]:

$$\begin{aligned} Q = \|\mathbf{e}_{new}\|^2 = \mathbf{x}_{new}\left(\mathbf{I} - \mathbf{P}_h\mathbf{P}_h^T\right)\mathbf{x}_{new}^T = \\ \sum_{i=1}^{K} \mathbf{x}_{new}\left(\mathbf{I} - \mathbf{P}_h\mathbf{P}_h^T\right)\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\mathbf{x}_{new}^T = \sum_{i=1}^{K} q_i \end{aligned} \tag{5.7}$$

where $\mathbf{I}$ is the identity matrix and $\boldsymbol{\xi}_i$ is the ith column of the identity matrix. Index i and K indicate a specific feature and the total number of measured features, respectively. The contribution of the measured value in feature i to Q equals q_i. High values indicate abnormal behaviour of this particular feature. However, the contribution of each feature in the NOC data should be taken into account as well: a large contribution of a feature becomes less meaningful if such contribution values were also observed for the NOC samples. Therefore, all contributions were studied relatively to the variance of the NOC residuals $\mathbf{E}_h$ [15]:

$$rq_i = \frac{q_i}{\tilde{\Lambda}_{i,i}} \tag{5.8}$$

Where $\tilde{\Lambda}_{i,i}$ indicates the element in the ith row and column of $\tilde{\mathbf{\Lambda}}$. The set of relative contributions for all features will be referred to as a personal health profile or personal biomarker.

5.3 Methods

5.3.1 Data

To assess the value of SHM for disease diagnosis, a set of urine samples of 193 healthy children and a set of 24 patients was measured using proton NMR spectroscopy. Eighteen patients were known to suffer from one of seven different IEM. For the other six patients, no IEM was diagnosed, but their metabotypes contained commonly prescribed drugs such as depakine and sabril. More details regarding the healthy and patient samples are specified in tables 5.1 and 5.2, respectively. Note that a subject had to be between 4 – 12 years old to participate in this study and be of Dutch ancestry. An equal amount of males and females were selected. No other selection criteria such as lifestyle and diet were imposed.

Table 5.1 Abnormal compounds present in urine 1D ^{1}H-NMR spectra from the healthy individuals. The metabolites were identified by comparison of the abnormal resonances to a database of NMR spectra of model compounds [5]. In cases where the overlap of resonances in the 1D spectrum was quite severe, 2D COSY NMR experiments were used to provide additional information and confirm that the metabolite identification based on the 1D spectrum was correct.

Compound	CS (ppm); multiplicity	Origin
Acesulfame	2.11d; 5.67q	Artificial sweetener
Acetaminophen	2.15s; 6.90d; 7.25d	Paracetamol
A-glucuronide*	2.16s; 5.12d; 7.13d; 7.34d	Paracetamol
A-*L*-cysteinyl*	2.15s; 6.99d; 7.51d	Paracetamol
A-*N*-acetyl-L-cysteinyl*	1.84s; 2.14s; 6.93d; 7.42d	Paracetamol
A-Sulphate*	2.17s; 7.45d; 7.31d	Paracetamol
Cyclamate	1.53 – 2.06m	Artificial sweetener
Mannitol	3.6-3.8v	Sweetener
N-Methylhydantoin	2.92s; 4.08s	Bacteria
TMA-oxide	3.54s	Fish meal

*Spectrum not completed interpreted; s=singlet; d=doublet; t=triplet; q=quartet; m=multiplet; v=various multipets
A = Acetaminophen; TMA = Trimethylamine

Table 5.2 Abnormal compounds present in urine 1D ^{1}H-NMR spectra from the patients. The metabolites were identified by comparison of the abnormal resonances to a database of NMR spectra of model compounds [5]. In cases where the overlap of resonances in the 1D spectrum was quite severe, 2D COSY NMR experiments were used to provide additional information and confirm that the metabolite identification based on the 1D spectrum was correct.

Compound	CS (ppm); multiplicity	Origin
Arginine	1.69m; 1.92m; 3.24q; 3.85t	Cystinuria
4-Amino-5-hexenoic acid	1.94m; 2.08m; 2.46m; 5.47m; 5.80m	Medication: Sabril
Dihydroxycholenic acid	0.67s; 0.80-0.94v	3β-Hydroxy-Δ^5-C_{27}-steroid dehydrogenase deficiency
Formiminoglutamic acid	2.00-2.22v; 2.47t	Formiminotransferase deficiency
Homogentisic acid	3.64s; 6.78m	Alkaptonuria
Hydantoin-5-propionic acid	2.00-2.22v; 2.51t	Formiminotransferase deficiency
3-Hydroxyisovaleric acid	1.33s; 2.55s	Isovaleric aciduria 3MCC-deficiency
Isovalerylglycine	0.94d; 2.02m; 2.18d; 3.94d	Isovaleric aciduria
Lysine	1.47m; 1.72m; 1.92m; 3.01t; 3.77t	Cystinuria
3-Methylcrotonylglycine	1.86d; 2.03d; 3.97d; 5.78m	3MCC-deficiency
2-oxo-1-pyrrolidine acetamide	2.10m; 2.48t; 3.52t; 4.01s	Medication: Piracetam
5-Oxoproline	2.20m; 2.43m; 2.55m; 4.36m	5-Oxoprolinuria
Taurine	3.27t; 3.43t(wide due to exchange)	Unknown; possibly nutrition
Trihydroxycholenic acid	0.73s; 0.80-0.94v	3β-Hydroxy-Δ^5-C_{27}-steroid dehydrogenase deficiency
Valproic acid	0.88t; 1.30m; 1.50m; 2.44m	Medication: Depakine

*Spectrum not completed interpreted; s=singlet; d=doublet; t=triplet; q=quartet; m=multiplet; v=various multipets
3MCC=3-Methylcrotonyl CoA carboxylase

The urine samples were centrifuged before analysis. A volume of 70 µl of a 20.2 mmol/l trimethylsilyl-2,2,3,3-tetradeuteriumpropionic acid (TSP, sodium salt; Aldrich) $^{2}H_2O$ solution was added to 700 µl of urine as a chemical shift reference (δ = 0.00) and as a lock signal. The pH of the urine was adjusted to 2.50 ± 0.05 with HCl. Finally, 650 µl of the sample was placed into a 5-mm NMR tube (Wilmad Royal Imperial; Wilmad LabGlass, USA).

^{1}H NMR spectra were obtained using a Bruker 500MHz spectrometer (pulse angle 90°, delay time 4s, number of scans 256, temperature 298K). The water resonance was suppressed by gated irradiation centred on the water frequency. Shimming was performed automatically on each sample prior to the data acquisition using the TopShim method from Bruker BioSpin. The phase and baseline were corrected manually.

The regions 0.2 – 4.7 ppm and 5.0 – 10.0 ppm were selected for further analysis in Matlab 7.14 (Mathworks, Natick, Massachusetts, U.S.A.). Next, the urine profiles were normalized

to the creatinine signal to correct for dilution effects. Equidistant binning with a bin size of 0.04 ppm was used to reduce the dimension of the normalized data from 30888 to 246 variables. Finally, the data was centred to zero mean and scaled to unit variance.

For some samples (see below), 2D COSY NMR spectra were also recorded for extra spectral information. The spectra were recorded at 500 MHz using 4k data points in F2 and a spectral with of 6002 Hz. For all samples, 256 increments and 16 scans per increment were used. The TR was 2s, during which the water resonance was presaturated. Prior to Fourier transformation, a since function was applied in both time domains.

5.3.2 Ethics statement

The medical ethical committee of the Radboud University Medical Centre in Nijmegen, The Netherlands, approved the study protocol. Informed verbal parental consent was obtained for all volunteers. At that time consultation of the medical ethical committee of the Radboud University Medical Centre in Nijmegen, The Netherlands resulted in the advice to ask the collaboration of the parents/caretakers of the children via an information letter explicitly stating that their collaboration was on a voluntary basis and that all samples would be fully anonymized and that all samples would be destroyed at the latest 5 years after the sample collection.

5.3.3 SHM model construction and validation

First, the raw NMR spectra of the 193 healthy individuals were extensively screened by an experienced clinical practitioner to rule out any abnormal metabolic patterns in these samples. Seventeen samples with abnormal patterns related to dietary influences and drug intake were identified. These samples were marked as abnormal and used to validate the SHM approach since detection of abnormal patterns due to diet and drugs is in principle no different from detection of abnormalities related to a disease. Additionally, the set of 24 patients was used for validation. In all samples, the abnormal metabolites were identified by the clinical practitioner by comparison of the abnormal resonances to a database of NMR spectra of model compounds [5]. In cases where the overlap of resonances in the 1D spectra was quite severe, 2D COSY NMR experiments were used to provide additional information and confirm that the metabolite identification based on the 1D spectra was correct.

The SHM model was constructed on the basis of 120 training samples that were selected from the remaining set of 176 binned ^{1}H NMR spectra of normal (healthy) individuals by the Kennard Stone algorithm [16]. The optimal number of factors in the PCA model was determined by a bootstrapping algorithm called NUMFACT [17]. In essence, the PCA

factors determined for each resampling were compared for changes. Factors that changed significantly from one resampling to the next were probably due to noise and excluded from the model.

Validation of identification of abnormal metabolites was performed by applying the left-out 56 healthy; 17 healthy, but abnormal; and 24 patient samples to the SHM model. An imposed significance limit (α) of 5% was used. Note that centring and scaling of the test data was based on the feature means and standard deviations of the training data.

All analysis was performed using in-house developed algorithms. Bootstrapping to estimate the number of factors in PCA was performed with PLS_Toolbox 6.7.1 [18].

5.4 Results

5.4.1 Inspection of the ^{1}H NMR data: current clinical practice

The data was first analysed according to current clinical practice, namely by visual inspection and by means of PCA score plots. This inspection was required to select the NOC samples (healthy individuals) on which the SHM model could be trained.

A clinical expert visually inspected the NMR data of the 193 healthy children and the 24 patients. Ten exogenous metabolites related to diet or drug intake were observed in the set of 193 children. More details are specified in table 5.1. Seventeen samples contained metabolites related to bacteria, a fish meal, paracetamol, or cyclamate. The artificial sweeteners Acesulfame K and mannitol were present in the metabotype of such a large number of healthy individuals that they were not marked as abnormal metabolites. Fifteen abnormal metabolites were observed in the 24 patient samples. These were related to 7 IEM (18 patients) as well as medication (6 patients). More details are specified in table 5.2.

In figure 5.1, a PCA score plot of all samples is shown (autoscaled data). The samples were coloured according to the observations made by the clinical expert. Clearly, many abnormal metabotypes could not be distinguished from healthy samples this way. Alternative colourings of the plot indicated no trends related to age, gender or other demographics either.

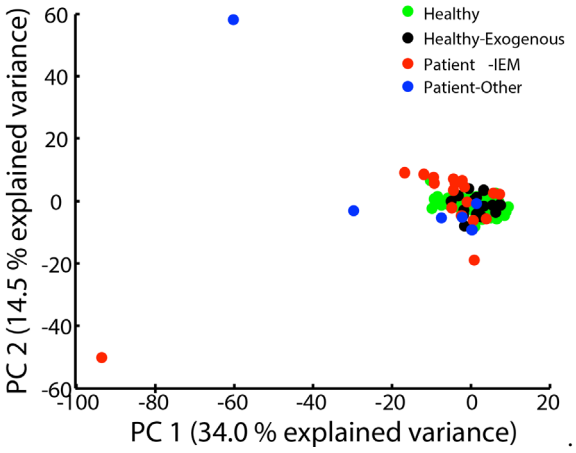

Figure 5.1 PCA score plot of the autoscaled data. Note that the points are coloured according to the observations made by a clinical expert as healthy; healthy, but exogenous metabolites were present; patients diagnosed with IEM; and other patients.

5.4.2 Statistical health monitoring

The SHM model was constructed on the basis of 120 healthy metabotypes. The clinical expert had not detected any of the exogenous metabolites listed in table 5.1 in these samples, except for the articial sweeteners Acesulfame and mannitol. This means that future samples that contain exogenous metabolites related to fish, paracetamol intake, etc will be marked as abnormal by the model even if they are healthy. This can be undesirable if the sole purpose of the SHM model is disease diagnosis. We will further elaborate upon this choice of NOC samples in the discussion

Eighty-three percent of the total variation in the NOC data was estimated to be systematic by NUMFACT. This variation was modelled by the first 16 factors. Next, the metabotype of the validation samples was automatically inspected using the constructed SHM model. As shown in figure 5.2, Q-values of the abnormal metabotypes were clearly larger compared to the normal metabotypes. Using the imposed significance limit of 5%, all normal and abnormal metabotypes were correctly identified. Note that the cut-off point $Q_{5\%}$ to mark a patient's metabotype as abnormal was completely based on the training samples (equation 5.4).

The 17% variance left out of the model must correspond to individual variations, which did not necessarily belong to the NOC of the whole group. This unexplained variance partly re-

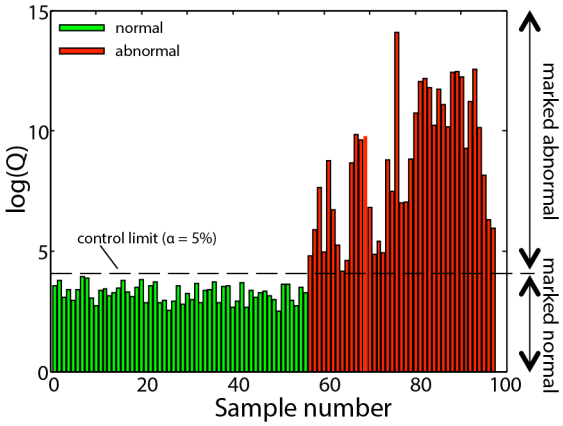

Figure 5.2 Q-values obtained by statistical health monitoring for 56 healthy and 42 abnormal metabotypes.

appeared as Q contribution. Therefore, metabotypes of normal individuals did not have zero contribution for every feature. Statistically speaking, with the chosen significance level 5% of the samples that are within NOC are expected to be incorrectly marked as abnormal (i.e. 3 individuals). In this case all normal individuals were correctly detected which is related to size of our test cohort.

For metabotypes marked as abnormal, the abnormal metabolites were identified via the relative feature contributions to Q. The set of relative contributions can be considered as a personal biomarker for that individual since they highlight how and how much this individual is different from NOC. The contribution can be visualized in a so-called contribution plot, which is commonly done in industrial process control, or in the original NMR spectrum to integrate SHM in current clinical practice. In a contribution plot the relative contribution is plotted against the chemical shift value. Three examples are presented in figures 5.3a, c, and e. In each figure, high peaks relative to the baseline indicate resonances that were abnormal with respect to NOC. An advantage of contribution plots is that features with a large contribution are easy to identify, even if they have a low intensity in the original NMR spectrum (e.g. the resonances between 9.6 – 9.8 ppm in figures 5.3e and 5.3f). In contrast to contribution plots, visualization in the NMR spectrum itself allows the user to make combined use of contribution values as well as NMR knowledge such as multiplet structure to make a diagnosis. As shown in figures 5.3b, d, and f, the contribution values are colour coded in this type of visualization.

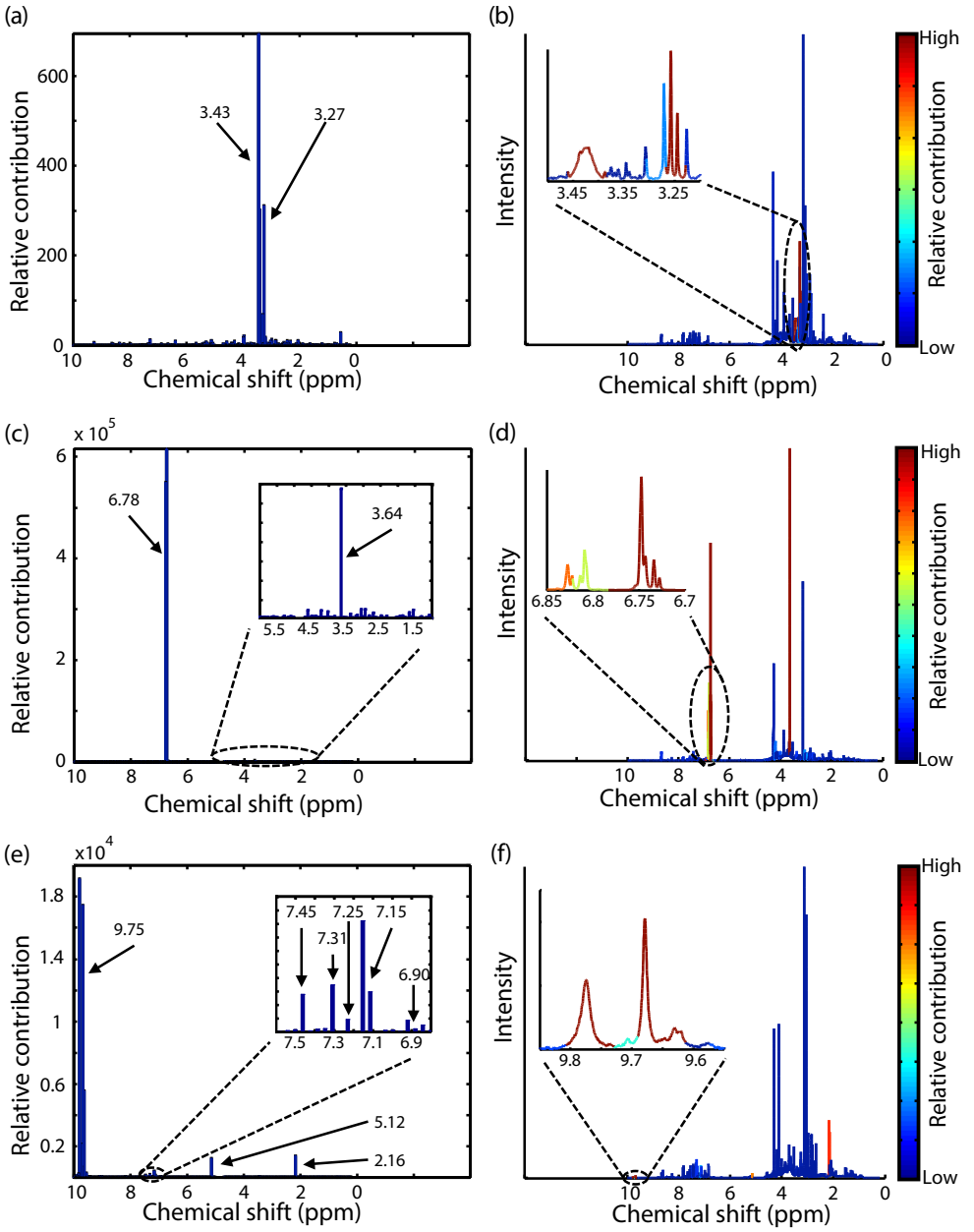

Figure 5.3 Feature contributions visualized in a contribution plot and the original NMR spectra for three abnormal metabotypes. The abnormal metabolites are related to **(a, b)** high concentrations of taurine, **(c, d)** alkaptonuria, and **(e, f)** paracetamol comsumption. The arrows indicate the resonance corresponding to the middle of a bin. Each bin had a width of 0.04 ppm.

The abnormal metabotypes were further inspected using both representations of the relative contribution. All IEM, were correctly diagnosed this way. Similarly, all abnormalities related to diet and the different types of medication were correctly identified. Most IEM were diagnosed via key resonances – biomarkers relating to the specific IEM; not always were all relevant biomarker resonances for a specific IEM marked as abnormal. This is similar to visual inspection of the data: resonances with a high degree of splitting or overlap cannot be discerned from noise in 1D spectra.

In order to illustrate the principles of relative contributions and the procedure to follow to establish a potential diagnostic better, we will describe the three examples in figure 5.3 in more detail below. The complexity of these examples is progressively rising in terms of spectral interpretation meaning that correct identification by SHM is more challenging.

Case 1

The first example is considered relatively simple because only two resonances are involved. As shown in figure 5.3a, SHM clearly marked two resonances at 3.27 and 3.43 ppm as abnormal: the relative contribution was much larger compared to the contribution of other resonances. Similar to visual inspection of the data, the metabolite corresponding to these abnormal resonances was identified by comparison of the resonances to a database of NMR spectra of model compounds. These particular two resonances correspond to taurine indicating that the metabotype of this patient contained abnormally high concentrations of taurine. This was confirmed by visual inspection of the spectrum by the clinical expert. At the moment the cause of the high concentrations of taurine in the metabotype of this patient is unknown. Perhaps, the abnormality can be related to diet, e.g. consumption of energy drinks.

Case 2

The contribution plot of this patient also showed two resonances that were abnormal: a singlet at 3.64 ppm and a multiplet at 6.78 ppm (figures 5.3c and 5.3d). However, inspection of this plot was more difficult compared to case 1. The multiplet was much easier to detect compared to the singlet because the NMR spectra of healthy individuals did not contain much signal around 6.78 ppm. Therefore, the relative contributions of this multiplet were very large. In contrast, the singlet at 3.64 ppm was positioned in a crowded region of the spectra resulting in much lower contribution values. However, compared to the contributions at the surrounding chemical shifts, the singlet at 3.64 was still clearly abnormal. This shows that inspection of contribution plots should not only focus on the absolute value of the contributions, but on their size relative to the contribution that is

observed for most chemical shifts. One could say that for each individual the "Q baseline" must be used to determine if a particular peak is abnormal or not. The abnormal singlet at 3.64ppm and the multiplet at 6.78 ppm indicated that the metabotype of this individual contained a large concentration of homogentisic acid [5]. Thanks to this, the patient was diagnosed with the IEM alkaptonuria. Alkaptonuria is caused by a deficiency of the enzyme homogentisic acid oxidase in tyrosine catabolism[5]. This results in high concentrations of homogentisic acid in the urine of such a patient.

Case 3

As shown in figure 5.3e, eight regions in the NMR spectrum of this individual had abnormal contributions. Similar to case 2, the contributions of the relevant resonances again differed by orders of magnitude.

Comparison of the abnormal regions to spectra of model compounds clearly indicated that the metabolites acetaminophen, acetaminophen-glucuronide, and acetaminophen-sulphate were present in high concentrations (see table 5.1). This is caused by consumption of paracetamol by this individual. As shown in table 5.1, this drug can actually be detected in urine via abnormal concentrations of five metabolites. A number of these metabolites will be visible in the NMR spectrum depending on how the drug was metabolized. In this case no high Q-values were observed at resonances 1.84, 6.99, and 7.51 ppm. This indicates that the compounds A-*N*-acetyl-*L*-cysteinyl and A-*L*-cysteinyl were either present in very low concentration, or that in this particular case paracetamol was not metabolised into these metabolites. This was confirmed by visual inspection of the NMR spectrum. Due to the large number of resonances involved, identification of paracetamol intake via SHM is considered more difficult compared to the previous two cases. Additionally, the intensities of the resonances involved are much lower which makes diagnosis even more difficult.

Note that for all individuals who consumed paracetamol, the resonances around 9.8 ppm were also marked as abnormal. These resonances have not been described in literature. However, by means of a simulated NMR spectrum of paracetamol in the Bruker software we ascribe these resonances to NH-groups in the molecule.

5.5 Discussion

In this study, SHM was introduced as a valuable tool for diagnosis of a multitude of possible (rare) diseases. The method was successfully applied in a case study involving diagnosis of several IEM as well as metabolic abnormalities related to drug consumption and diet.

First, the metabotype of each individual was marked as normal or abnormal: 100% of the "healthy" and 100% of the abnormal metabotypes were correctly identified. Next, it was shown that feature contributions can be used to identify the abnormal metabolites. The contributions are very easy to calculate without prior knowledge. Prior knowledge, however, is required to interpret them and relate the abnormal features to a disease. Therefore, SHM should be regarded as a decision support tool for diagnosis. In case of rare diseases, SHM is the only tool available to detect the abnormalities. In case of more common diseases, the SHM-based metabotype screening could be followed by more classical targeted approaches (e.g. a two-class classifier) to confirm the diagnosis.

The first step of SHM, detection of abnormal metabotypes in a multivariate fashion, is functioning in a reliable way. The second step concerns identification of the abnormality. This identification relies on a univariate evaluation of the individual contributions of each feature or measured signal to Q. These contributions should be studied with some caution due to the so-called smearing effect: contributions from abnormal features can propagate to other features meaning that fault free features can show increased contribution [15]. This a well-known issue in industrial process control that has been greatly discussed in literature [13]. The smearing effect is a direct consequence of the fact that an SHM model is constructed on the basis of normal metabotypes. Because of this the model is very well able to detect when a metabotype is abnormal. However, when the abnormal metabolites are identified via the Q-statistic again information from the normal (NOC) individuals has to be used (see equation 5.7). The model assumes that the correlations between metabolites in the abnormal metabotype are the same as those encountered in the NOC samples. This doesn't have to be the case. This imperfect assumption may lead to some false positives i.e. some metabolites can be marked as abnormal while they are not. Unfortunately, the smearing effect cannot be avoided. In this study, the smearing effect was minimized by using partial decomposition of the Q-statistic, instead of the commonly used complete decomposition method. This ensures that the contribution of an abnormal feature will always be greater than the contribution given to the "good" feature [15].

Some works in the process control literature suggest the use of control limits for determining the significant feature contributions. However, this should be discouraged since the (biological) unrelated features might also show an increased contribution due to the smearing effect [15]. Therefore, we prefer to rely on human expertise by visually inspecting the contribution plots.

Selection of NOC samples is a critical step in the construction of an SHM model. The proposed method detects deviations from these normal samples. This deviation can be

related to disease, but also a healthy sample with a deviation that is not present in the NOC set – e.g. the paracetamol example presented in case study 3. When disease diagnosis is the goal of the SHM model, examples such as case 3 are false positives. As shown in figure 5.1, the patient samples differed more from the NOC samples (the healthy group) compared to the group of healthy samples that contained exogenous metabolites. This was the main reason why the latter group was not included in NOC: we wanted to investigate if SHM was also able to detect these smaller deviations from NOC. This group should be included in NOC, however, if the user only wants to detect metabolites related to disease. Therefore, we also briefly investigated this disease diagnosis model. A direct consequence of the fact that the NOC now contained extra inter-individual variation due to diet and medication was that the cut-off value for Q (equation 5.4) was increased. This means that samples will less quickly be marked as abnormal, increasing the chance of false negatives. In this case, however, all patient samples were still correctly diagnosed. In contrast to the SHM model presented in the results section, all samples with metabolites related to medication and drugs were now marked as normal. In this feasibility study, the NOC set was a small population of healthy children. These samples matched in age and ancestry to the expected IEM patients. No restrictions on factors such as lifestyle were imposed to ensure enough diversity within the NOC set so that it is representative of future patient samples. However, due to the size of our cohort most likely not all possible factors such as all types of medication were included in NOC. We expect that a much larger cohort of randomly selected NOC samples would contain most of the common diets; types of medication; and other factors, thereby avoiding the occurrence of false positives related to this. Additionally, if false positives occur later on, the NOC set can be updated with these samples. Note that the cohort should match the expected demographics of the patients are closely as possible since the larger the biological variation of the NOC samples the more difficult it will be to detect a subtle abnormality related to disease.

When working with large cohorts of NOC samples, an interesting research line would be to see if sub-populations of normal individuals with completely different characteristics due to e.g. environmental factors can be identified. Each sub-population has its own NOC. In such a case, a SIMCA-like model structure where a separate SHM model is constructed for each population seems more appropriate compared to one general model that was used in the present study. Matching new samples only against NOC of the correct sub-group could greatly enhance the power of the SHM model for disease diagnosis. If the subpopulations are unknown, perhaps a clustering approach such as mixture modeling can be used to define them.

Three additional future development lines can be defined for SHM: (1) connection of SHM output to a disease database, (2) development of personalized health control, and (3) application of SHM in clinical trials. The first research line could be implemented in SHM in the form of a database of known disease, which would automatically link the abnormalities detected with a list of potential diseases.

The second perspective is to define NOC at an individual level instead of a population or sub-group based one. To do so, longitudinal metabotyping experiments should be performed. The SHM model would then very precisely describe the NOC metabotype profile because no intra-individual differences have to be taken into account. In consequence, SHM would be able to detect more subtle abnormalities. An additional advantage of longitudinal studies for detection of abnormal metabotypes is that the user can accumulate information from a series of measurements. One could check whether measurements appear randomly distributed between the control limits or if a structure is appearing, signalling the start of a deviation from NOC. Such tests may greatly improve the power of SHM for disease detection. Identification of the specific abnormality may be improved by studying contributions relative to the last k timepoints instead of all NOC samples. The abnormal metabotype should be most similar to the last metabotypes that were measured before the individual became ill.

In this study, NMR was used to assess the metabotype of each individual because it is a very stable technique with a detection limit in the low micromolar range. This technique has been used routinely for over 20 years in Radboud University Medical Centre in Nijmegen to diagnose IEM. Although NMR is a valuable analytical platform to diagnose IEM, it is not necessarily the best technique for other diseases. Other data types such as results of classical blood tests or more advanced measurements such as 2D-NMR and LC-MS should be used if they are known to provide more relevant information. In principle, SHM can be applied to any type of data. For each application it is important to take into account the structure of the data and adjust the model accordingly. Here, PCA was used to describe healthy metabotypes. Multiway data coming from 2D-NMR or LC-MS could be evaluated using a PARAFAC or Tucker3 structure [19]. Because SHM can be applied to any data type, it will most likely not only find application in metabolomics, but also in other fields such as proteomics or genomics.

5.6 Conclusion

Due to the complex nature of metabolomics data, multivariate statistics are required to interpret the data. Unfortunately, current multivariate tools can only diagnose diseases in a

targeted fashion; a separate model is required for each disease. Additionally, such tools are not always applicable to rare or orphan diseases. Abnormal metabotypes can sometimes be detected in an untargeted fashion by visual comparison of the data. However, detection of subtle abnormalities and abnormal patterns is extremely subjective and time-consuming. An alternative approach, SHM, was proposed in this study.

In SHM, the metabotype of an individual is compared to normal (healthy) metabotypes in a multivariate manner. Any abnormal patterns are indicated by the method. Subsequently, this information can be used for diagnosis. In this study, SHM was successfully applied for diagnosis of various metabotypic abnormalities related to diet, drug intake and IEM.

SHM is a general method that is not only applicable to metabolomics data. Additionally, the method offers perspectives in the framework of personalized health.

References

1. Nicholson, J.K., et al., *Metabolic phenotyping in clinical and surgical environments.* Nature, 2012. **491**(7424): p. 384-392.
2. Lindon, J.C., J.K. Nicholson, and E. Holmes, *The handbook of metabonomics and metabolomics*. 2007, Amsterdam, The Netherlands: Elsevier.
3. Madsen, R., T. Lundstedt, and J. Trygg, *Chemometrics in metabolomics—A review in human disease diagnosis.* Analytica Chimica Acta, 2010. **659**(1–2): p. 23-33.
4. Holmes, E., I.D. Wilson, and J.K. Nicholson, *Metabolic Phenotyping in Health and Disease.* Cell, 2008. **134**(5): p. 714-717.
5. Engelke, U.F.H., et al., *Handbook of 1H-NMR spectroscopy in inborn errors of metabolism: body fluid NMR spectroscopy and in vivo MR spectroscopy*. 2007: Heilbronn: SPS Verlagsgesellschaft.
6. Lindon, J.C., E. Holmes, and J.K. Nicholson, *Metabonomics in pharmaceutical R & D.* FEBS Journal, 2007. **274**(5): p. 1140-1151.
7. Dunn, W.B., N.J.C. Bailey, and H.E. Johnson, *Measuring the metabolome: current analytical technologies.* Analyst, 2005. **130**(5): p. 606-625.
8. Lenz, E.M. and I.D. Wilson, *Analytical Strategies in Metabonomics.* Journal of Proteome Research, 2006. **6**(2): p. 443-458.
9. Trygg, J., E. Holmes, and T. Lundstedt, *Chemometrics in Metabonomics.* Journal of Proteome Research, 2006. **6**(2): p. 469-479.
10. Smolinska, A., et al., *Interpretation and Visualization of Non-Linear Data Fusion in Kernel Space: Study on Metabolomic Characterization of Progression of Multiple Sclerosis.* PLoS ONE, 2012. **7**(6): p. e38163.
11. Chen, T., et al., *Random Forest in Clinical Metabolomics for Phenotypic Discrimination and Biomarker Selection.* Evidence-Based Complementary and Alternative Medicine, 2013. **2013**: p. 11.
12. Melnikova, I., *Rare diseases and orphan drugs.* Nat Rev Drug Discov, 2012. **11**(4): p. 267-268.
13. Joe Qin, S., *Statistical process monitoring: basics and beyond.* Journal of Chemometrics, 2003. **17**(8-9): p. 480-502.
14. Bersimis, S., S. Psarakis, and J. Panaretos, *Multivariate statistical process control charts: an overview.* Quality and Reliability Engineering International, 2007. **23**(5): p. 517-543.
15. Alcala, C.F. and S. Joe Qin, *Analysis and generalization of fault diagnosis methods for process monitoring.* Journal of Process Control, 2011. **21**(3): p. 322-330.
16. Kennard, R.W. and L.A. Stone, *Computer Aided Design of Experiments.* Technometrics, 1969. **11**(1): p. 137-148.
17. Henry, R.C., E.S. Park, and C.H. Spiegelman, *Comparing a new algorithm with the classic methods for estimating the number of factors.* Chemometrics and Intelligent Laboratory Systems, 1999. **48**(1): p. 91-97.
18. Wise, B.M., *PLS Toolbox for use With MATLAB™*, 1992, CPAC: Washington.
19. Louwerse, D.J. and A.K. Smilde, *Multivariate statistical process control of batch processes based on three-way models.* Chemical Engineering Science, 2000. **55**(7): p. 1225-1235.

Appendices

A. Schematic depiction of SHM when monitoring two metabolites

In figure A.1, disease detection is illustrated for a problem where 2 metabolites are monitored. Note that in comparison to the main text in this chapter, a second statistic (T^2) is show here as well. This statistic is often used in statistical process control to monitor the values of the scores of the PCA model. In our application this statistic does not offer any additional information compared to the Q-statistic. However, this statistic can be useful when SHM is applied to other types of data.

Figure A.1 shows a 1-factor (or principal component) PCA model that describes the main differences between the NOC samples. Outlier 1 has a low value for the Q-statistic, but a high value for the T^2-statistic. The metabotype of this individual is explained well by the model but is different from the NOC metabotypes in the sense that it is an extreme object (it has an extreme score). This means that the sample has extreme metabolite concentrations, but the overall between-metabolite correlation pattern is conserved, e.g. due to improper correction for dilution effects. Outliers 2 and 3 have a high value for the Q-statistic. The metabotype of these individuals is not described well by the model because the relation between the two metabolite concentrations is very different from NOC. In other words, the Q-statistic can detect abnormal correlations between the metabolites. Therefore this statistic is the most suitable for disease diagnosis.

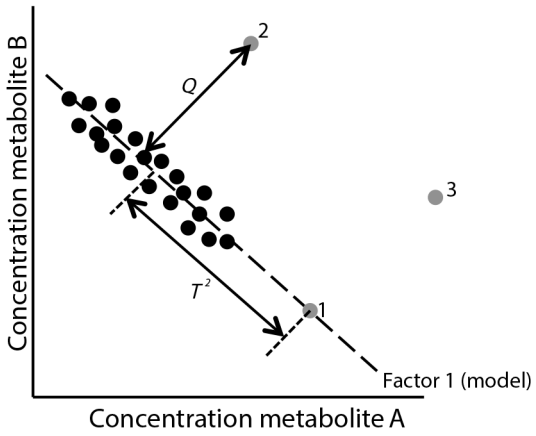

Figure A.1 Disease detection via statistical health monitoring for a two metabolites problem. The black dots indicate NOC samples. The grey dots are outliers.

6

A NOVEL APPROACH TO DIAGNOSIS AND FOLLOW-UP OF INDIVIDUAL PATIENTS BY SPARSE MODELING

The –omics technologies are becoming increasingly important in health care and are expected to contribute to personalized health care. Statistical health monitoring (SHM) was introduced in chapter 5 for analysis of -omics data to automatically identify the disease response in an individual patient. This approach could be of use in all sorts of applications such as diagnosis of rare diseases, analysis of individual patterns in disease manifestation, disease monitoring, or personalized therapy.

SHM essentially combines estimation of Mahalanobis distances (MD) with principal component analysis (PCA). It is well known that the dimension reduction step via PCA can hamper reliable identification of the disease response in a patient. Therefore, sparse SHM (sSHM) is introduced in this chapter. The method combines estimation of the MD with variable selection by inclusion of an ℓ_1-norm constraint.

Simulations are used to show that the sSHM model can identify the disease response in an individual patient more reliably compared to SHM. Subsequently sSHM is applied to urine ^{1}H NMR metabolomics data for diagnosis of several orphan diseases.

This chapter has been adapted from: Engel, J., Blanchet, L., Engelke, U. F. H., Wevers, R. A., & Buydens, L. M. C. A Novel Approach to Diagnosis and Follow-up of Individual Patients by Sparse Modeling. *Submitted*.

6.1 Introduction

The advent of the –omics technologies has allowed for analysis of DNA, proteins, metabolites and other molecules in a patient at an unprecedented level [1, 2]. Because of this, factors that influence health can be studied in great detail allowing for precise diagnostics [3]. Omics techniques are becoming increasingly important in health care and are expected to contribute to personalized health care. Metabolomics, for example, has been used to diagnose cancer states, diabetes, cardiovascular diseases, neurological diseases and inborn errors of metabolism (IEM) amongst others [4-8]. Many different types of samples can be used in –omics experiments ranging from tissues to body fluids. Sampling is becoming increasingly non-invasive focusing on easily obtainable body fluids such as urine or analysis of exhaled breath [9, 10].

Commonly, an –omics experiment involves hundreds to thousands of measured variables per patient. Therefore, analysis of the acquired data with (multivariate) statistical approaches is a crucial step to diagnose a disease or monitor the healthy state [8, 11]. Often classification models are used to classify groups of patients with a specific disease from healthy controls. These models focus on populations of patients with a specific disease. This might be impractical in a clinical setting for a number of reasons, including: (1) in the case of diagnosis of rare diseases typically too few samples are available for a useful comparison between patients and controls; (2) a separate model is required for each disease; and (3) unknown metabolic perturbations, i.e. caused by an unknown disease, may not be detected or falsely interpreted. Additionally, by definition population-based approaches are designed to be the best for the average population and not for specific individuals. Individuals are genetically diverse and have different exposures and response characteristics to disease [12]. These individual responses are hypothesized to overlap in large population models and may therefore be very challenging to detect [1]. Clearly, statistical modeling approaches beyond population-based models are required to improve analysis of –omics data for (personalized) health care.

In chapter 5, statistical health monitoring (SHM) was introduced as an alternative tool for disease diagnosis in metabolomics [4]. The method can be used to identify the disease response in an individual patient in an untargeted way. This information can be used for disease diagnosis, analysis of individual patterns in disease manifestation, disease monitoring, personalized therapy, etc.

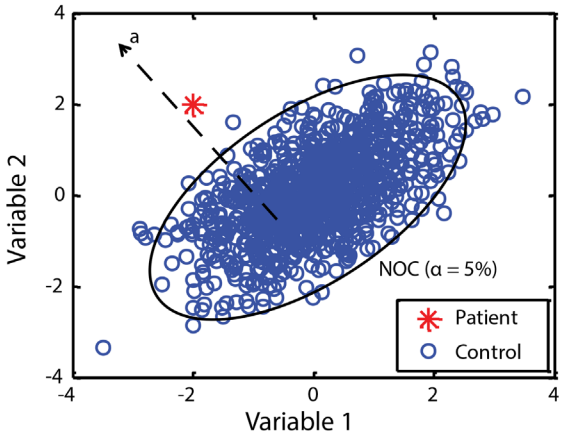

Figure 6.1. Comparison of a patient (marked red) to a set of controls (marked blue) by the Mahalanobis distance. Note that the solid black line corresponds to the 95% confidence sphere, i.e. the NOC. The dotted arrow indicates the direction of canonical variate **a**, which is used to determine on which variables the patient differed most from the NOC.

The basic concept of SHM is shown in figure 6.1. A set of control samples is used to define the so-called normal operating conditions (NOC) represented by the black ellipse in the figure. The NOC represents the expected characteristics of (a) healthy metabolic phenotype(s). The NOC can be defined for the population, or for the individual. When a patient sample falls outside the black ellipse it is marked as significantly different. In that case a second step is used to determine which pattern of variables in the patient sample is abnormal with respect to the NOC. This way, disease responses can be assessed in an individual manner. The fact that only data of healthy people is required to construct the SHM model is another advantage of this approach. Because of this property, it was possible to successfully apply SHM to urine NMR metabolomics data for diagnosis of a number of rare diseases, namely IEM [4].

The SHM model essentially uses the Mahalanobis distance (MD) to compare the patient to the NOC (see appendix E) [4]. However, it is well known that the MD cannot be estimated reliably when the number of control samples is smaller than the number of variables [13]. SHM circumvents these issues by projecting the data to a lower dimensional space using principal component analysis (PCA). A major drawback of the PCA step is, however, that it allows for interaction between the abnormal variables (i.e. variables on which the patient differs from NOC) and the normal variables [14, 15]. Because of this, reliable identification of the abnormal variables is challenging (normal variables may appear to be abnormal). This is sometimes referred to as the smearing effect and hampers disease diagnosis. Additionally,

selection of the correct number of principal components is a crucial step [16]. In our experience, standard selection approaches such as screeplots do not provide a clear solution for –omics data.

In this chapter we propose an alternative SHM method, that is, sparse SHM (sSHM). The key idea consists of combining the MD with a variable selection method instead of a dimension reduction method. The main advantage of this approach is that abnormal variables are more reliably identified this way as will be shown by simulation. Additionally, subtle differences between the patient and NOC can be detected this way while they could be masked by the accumulative noise effect of irrelevant variables when no variable selection is applied (i.e. when the MD is used). Variable selection in sSHM is achieved by inclusion of an ℓ_1-norm constraint during estimation of the MD. A fast algorithm is employed to compute the entire solution path of the model, i.e. it finds all solutions with $1,2,\cdots,p$ selected variables. This path can be graphically depicted and provides an intuitive way to identify those variables on which the patient differs from the NOC, i.e. the disease response of the patient. The sSHM method is applicable in cases where the number of control samples is smaller than the number of variables, and, unlike SHM, no parameter optimization is required.

The next section will outline the concept of sSHM and its mathematical background. In the subsequent sections the properties of sSHM are studied based on simulated data. This data is also used to compare the model to SHM and an alternative variable selection approach originating from the field of industrial process monitoring. Finally, a case study involving a real metabolomics data set for diagnosis of IEM is used to demonstrate the value of sSHM for disease diagnosis in practice.

6.2 Theory

In sSHM an individual patient sample is compared to the NOC in two steps. In the first step the patient data is matched against the NOC and marked as normal (healthy) or possibly abnormal (possibly ill). When a sample is marked as abnormal a second step is used to identify the abnormal variables, i.e. the disease response in this individual. Below, it is first described how the Mahalanobis distance can be used to compare a patient to the NOC in this way. Next, the sSHM model, which combines the MD-statistic with variable selection, is introduced. Due to variable selection more subtle differences between the patient and NOC can be detected in the first step. Additionally, the disease response can be better identified in the second step.

6.2.1 The Mahalanobis distance

To be able to compare a patient to NOC, a set of healthy control samples ($\mathbf{X}_\mathrm{h}$) that represent the NOC well has to be selected. Note that the choice of control samples defines what kind of patterns in the patient sample will be marked as abnormal. For example, if the patient data contains signal due to paracetamol intake, but the NOC data does not, this signal will be marked as abnormal. More details regarding the choice of NOC samples in the context of SHM are provided in chapter 5 [4]. We assume that the distribution of the NOC samples is multivariate normal.

First, the patient is compared to the NOC by the squared Mahalanobis distance (MD^2) [17]:

$$MD^2 = \left(\mathbf{x}_p - \boldsymbol{\mu}_h\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_p - \boldsymbol{\mu}_h\right)^T > c \tag{6.1}$$

where $\mathbf{x}_p$, $\boldsymbol{\mu}_h$, and $\boldsymbol{\Sigma}$ indicate the row vector of patient data, the mean vector of the control samples, and covariance matrix of the control samples, respectively. Geometrically, expression 6.1 tests whether the patient sample falls inside the confidence sphere of the control samples, where the upper limit c is traditionally derived from a scaled F-distribution [17]. Note that the sphere defines the NOC: any sample that falls inside the sphere is marked as normal. The shape and location of the NOC sphere are defined by $\boldsymbol{\mu}_h$ and $\boldsymbol{\Sigma}$. An example is shown in figure 6.1.

If a sample falls outside the NOC it is abnormal and must be further inspected. The MD can also be used for this step. For this purpose another expression of the MD is used. More specifically, it has been shown that the squared Mahalanobis distance is the weighted distance between the scalar projections of $\mathbf{x}_p$ and $\boldsymbol{\mu}_h$ onto the vector $\mathbf{a} = \left[\mathbf{x}_\mathrm{p} - \boldsymbol{\mu}_h\right]\boldsymbol{\Sigma}^{-1}$, i.e. $MD^2 = \left(\left[\mathbf{x}_\mathrm{p} - \boldsymbol{\mu}_h\right]\mathbf{a}^\mathrm{T}\right)^2 / var(\mathbf{X}_h\mathbf{a}^\mathrm{T})$ [18]. Vector $\mathbf{a}$ is a canonical variate (CV) known from techniques such as Fisher LDA and MANOVA. The CV is indicated by the dotted line in figure 6.1. It is a linear combination of the variables that maximizes the difference between $\mathbf{x}_\mathrm{p}$ and $\boldsymbol{\mu}_h$ relative to the variance of the controls [18]. Therefore, the coefficients of $\mathbf{a}$ can be studied to determine on which variables the patient sample differed most from NOC, where large absolute coefficients indicate abnormal variables. The abnormal variables are interpreted as the individual disease response of the patient. A clinical practitioner can use this information for example for disease diagnosis.

Typically, only a few of the measured variables are related to a disease. However, all measured variables are taken into account by MD^2. As shown in appendix A the effect caused by the disease in a few variables may be masked by the normal variation in the other

variables. This can result in a considerable loss in power and hampers identification of the abnormal variables when inspecting direction **a**, especially when the effect is small. Additionally, equation 6.1 cannot be applied when the number of control samples is smaller than the number of variables because the inverse of **Σ** cannot be computed. Because of this, a data-driven approach is introduced in this work to only take the most abnormal variables into account when comparing a patient to the NOC. This increases the power of the MD^2-test and improves identification of the abnormal variables.

6.2.2 Sparse Statistical Health Monitoring

The defining feature of our approach is that we first estimate the canonical variate **a** before the MD is computed. Additionally, we regularize estimation of direction **a** with an ℓ_1-norm constraint. This constraint is well known from techniques such as the LASSO and the elastic net [19]. It has the sparsity property in the sense that it will force some coefficients in **a** to be exactly zero indicating variables on which the patient was similar to the NOC. The MD is calculated using this sparse estimate to determine if the patient differs significantly from the NOC. In this sense the ℓ_1-norm constraint introduces a variable selection step since only the variables with a nonzero coefficient in **a** contribute to the estimate of the MD. Sparse estimation of **a** is very useful to identify the abnormal variables in a sample.

Sparse estimation of canonical vector a

To be able to combine estimation of canonical variate **a** with a constraint, we first note that **a** can be estimated by maximizing the Rayleigh quotient $[\mathbf{a}(\mathbf{d}^T\mathbf{d})\mathbf{a}^T]/[\mathbf{a}\boldsymbol{\Sigma}\mathbf{a}^T]$, where $\mathbf{d} = \mathbf{x}_p - \boldsymbol{\mu}_h$. Wu et al showed that maximizing the Rayleigh coefficient is equal to the following expression [20]:

$$\hat{\mathbf{a}} = argmin_{\mathbf{a}}(\mathbf{a}\boldsymbol{\Sigma}\mathbf{a}^T) \;\; s.t. (\mathbf{x}_p - \boldsymbol{\mu}_h)\mathbf{a}^T = 1 \tag{6.2}$$

Note that the inverse of **Σ** doesn't have to be estimated this way. However, the solution to **a** uses all variables and shares most of the problems as the normal MD. Therefore, an ℓ_1-norm constraint is placed on the direction **a** to obtain a sparse solution, i.e. incorporate variable selection:

$$\hat{\mathbf{a}} = argmin_{\mathbf{a}}(\mathbf{a}\boldsymbol{\Sigma}\mathbf{a}^T) \;\; s.t. (\mathbf{x}_p - \boldsymbol{\mu}_c)\mathbf{a}^T = 1, \sum_{i=1}^{p} |a_i| \leq \lambda \tag{6.3}$$

where a_i indicates the i-th coefficient of **a**, and λ is a fixed constant. The value of λ controls the amount of variables that are included in the model; when λ is small, most of the a_i will

be exactly zero. Expression 6.3 can be solved in different ways. Here we use a fast algorithm that was developed by Wu et al to solve a sparse LDA problem [20, 21]. Note that expression 6.3 can also be written as a penalized regression problem using a suitable weighting of the data. This way, standard algorithms for sparse regression such as the LASSO or elastic net can be used to estimate **a**.

Using a regularized estimate of **Σ** could provide some further improvement; also it stabilizes the algorithm used to solve expression 6.3. In this work we regularize **Σ** by adding the value $\delta * \log(p)/n$ to its diagonal elements, where δ is a small constant such as 2 or 4, p corresponds to the number of variables in the data and n to the number of training samples in matrix $\mathbf{X}_h$ [20]. More advanced methods are available such as linear or non-linear shrinkage methods [22]. However, these methods are computationally more demanding; additionally in this work we don't want to entangle estimating covariance matrices with variable selection.

Comparison of a patient to the NOC

In sSHM, the dissimilarity between the patient and the NOC is evaluated by the squared MD using a sparse estimate of **a**:

$$MD_i^2 = \left([\mathbf{x}_\mathrm{p} - \boldsymbol{\mu}_h]\mathbf{a}_{\lambda_i}^\mathrm{T}\right)^2 / var(\mathbf{X}_h\mathbf{a}_{\lambda_i}^\mathrm{T}) \tag{6.4}$$

where MD_i^2 indicates a sparse squared Mahalanobis distance based on i selected variables. As mentioned above, the power of the method is increased by variable selection. The variables are selected in a data-driven fashion by equation 6.3, where $\mathbf{a}_{\lambda_i}$ is its solution with $\lambda = \lambda_i$ chosen such that i variables are selected. Typically, a range of constants $\lambda_{i1} < \lambda_{i2} < \cdots < \lambda_{ik}$ exists such that equation 6.3 gives a solution with i nonzero coefficients (i selected variables). The value for λ_i in (6.4) is always set to the highest constant in this sequence, i.e. $\lambda_i = \lambda_{ik}$.

For each patient there is an optimal number of selected variables such that a possible difference with the NOC can be best observed. This number depends on how much the patient differs from the NOC in the abnormal variables with respect to the accumulating noise due to normal variation in the other variables (see appendix A). In practice, however, this number is unknown. In this work we use two approaches to resolve this issue. The first approach was suggested by Wang et al in the context of industrial process monitoring where they used domain knowledge to restrict the solution of their sparse model to i variables [23]. In the context of sSHM this means that the clinical practitioner must have a

rough idea regarding how many variables are affected by e.g. a disease in the patient data. Our simulation studies show that often a significant difference from NOC can be detected by selecting roughly this expected number of variables. The second approach is to determine i from the data by selecting that number of variables for which the largest distance MD_i^2 is observed. However, this is not a straightforward task since distances based on different amounts of selected variables are not directly comparable [24]. Therefore, a normalization step is used to fairly compare the MD_i^2-values:

$$MD_{\max}^2 = \max_{i=1\cdots p} \frac{MD_i^2 - E(MD_i^2)}{\sqrt{Var(MD_i^2)}} \tag{6.5}$$

where $E(MD_i^2)$ and $Var(MD_i^2)$ are the mean and variance of MD_i^2 for control samples, respectively. The values for $E(MD_i^2)$ and $Var(MD_i^2)$ are estimated from the control data by means of leave-one-out cross-validation (LOO-CV). Zou et al and Capizzi *et al.* also used this approach to select the optimal number of variables in their sparse industrial process control methods [25, 26].

After selecting a specific number of variables a significance test is required to determine if the patient indeed differs from NOC. For this purpose, the upper limit of MD_i^2 and $MD_{\max}^2$ is estimated from a generalized extreme value distribution. Note that the significance test does not test whether the sample is significantly different from the NOC on the specific variables selected by equation 6.3 for that sample. The fact that each sample (control and patient) may differ from NOC on completely different variables is taken into account. More details are provided in appendix B.

Identification of abnormal variables

As described above, the first step of sSHM is to use equation 6.3 to detect a significant difference between the patient and the controls. The second step involves interpretation: if a sample is found to be significantly different from NOC the subsequent goal is to identify most (*all*) abnormal variables, e.g. the individual disease response. Again, variable selection is useful in this respect. In principle, the abnormal variables should be selected first by the model and can thereby be identified this way. However, it is unclear how many abnormal variables are present in a sample and should therefore be selected. Expression 6.5 seems to be useful to automatically select all relevant variables. However, our simulations show that this test (that is designed to best observe differences between a patient and NOC) does not always reliably identify the abnormal variables (see e.g. figure 6.6a). Similarly, when the coefficients of $\mathbf{a}_{\lambda_i}$ for a specific number of selected variables are inspected it can be that too many noise variables are included masking the relevant variables (see e.g. figure 6.3a), or

that not all relevant variables are selected. Additionally, interpretation of a specific solution depends on scaling of the data (e.g. standardized or unstandardized coefficients) although the data is typically autoscaled to ensure that the ℓ_1-norm constraint 6.3 affects each variable equally. Therefore, we propose to inspect multiple solutions with different numbers of selected variables together in the so-called solution path figure to identify most of the abnormal variables. Examples will be provided in the results section in figures 6.4 and 6.6.

The solution path offers an intuitive way to see in which order the variables were selected by sSHM. Based on this information the variables can be ranked, where it is assumed that the most abnormal variables are selected first. The ranking is scale invariant. A clinical practitioner can inspect the top ranked variables to determine the patient's response to e.g. a disease, where the number of variables to inspect can be chosen based on practical considerations (i.e. it is not feasible to inspect more than 50 variables). Additionally, as soon as it is clear that a number of variables corresponding to a specific diagnostic marker have been selected, the solution path offers an intuitive way to quickly assess the selection rank of other variables that are biologically related to this marker.

6.3 Methods

We use simulated and real data to investigate the properties of the sSHM model under different structures of the NOC. The simulation study was also used to compare sSHM to competitive methods such as SHM with a focus on identification of the abnormal variables in a patient. These methods were applied according to the protocols described in the original papers [4, 26, 27]. The methods are also briefly explained in appendix E. In the main text of this paper, we mainly discuss the comparison between sSHM and SHM in the context of reliable identification of abnormal variables.

6.3.1 Simulation design

A simulation study was used to compare the SHM and sSHM models. Identification of abnormal variables by SHM is reliable when only one variable is abnormal [14, 15]. In this simulation, two variables were abnormal. Correct identification by SHM is not guaranteed in this case due to variable smearing (see appendix E) [14, 15, 27]. Therefore, this simulation design was enough to highlight the main difference between SHM and sSHM. The simulation was carried out for different structures of the NOC. Throughout the simulation it was assumed that the distribution of healthy controls was multivariate normal $N(\mathbf{0}, \mathbf{\Sigma})$. The covariance matrix $\mathbf{\Sigma}$ was constructed by multiplication of a predefined correlation structure $\mathbf{R}$ with variance values drawn from the uniform distribution $U(0.1,16)$.

It is difficult to simulate correlation structures **R** that closely resemble the complex structures of metabolomics data. Therefore, we studied three "simple" structures to be able to systematically explore the properties of the sSHM method. Note that the simulated data contained some aspects of real metabolomics data such as correlations between variables, grouping of variables, and irrelevant noise variables. Additionally, we applied sSHM to real NMR metabolomics data to study its performance for more complicated correlations (see section 6.3.2 below). The following correlation structures were considered in the simulation study:

1. **R1** Common correlation: all variables were correlated to each other with value ρ
2. **R2** Block-diagonal: blocks of variables were correlated to each other with value ρ. The different blocks were uncorrelated. Four blocks were defined; each block contained 25 variables. The first 25 variables corresponded to the first block, the 26th to 50th variables to the second block, etc.
3. **R3** Toeplitz structure: variables close to each other were more highly correlated compared to variables that were far apart. The distance between variables with indices $p1$ and $p2$ was defined as $|p1 - p2|$. For example, the distance between the first and fourth variable in the data is equal to three. The correlation between variables was given by $\rho^{|p1-p2|}$.

The correlation ρ was set to 0, 0.4, 0.6, and 0.8. The correlation structures are visualized in figure 6.2. Note that when the correlation was set to zero an independent covariance structure was obtained.

In the simulation the number of variables in the data set was varied between 10 and 100. Additionally, the number of control samples that was used to train the sSHM model (i.e. used to estimate $\boldsymbol{\mu}_h$and $\boldsymbol{\Sigma}$ in equation 6.3) was 50 or 1000. These samples were drawn from $N(\mathbf{0}, \boldsymbol{\Sigma})$ with a specific covariance matrix as defined above. Next, test samples were simulated. These samples were different from NOC on two variables. The amount of difference was varied by constant f, which ranged from 0 to 3 in steps of 0.25. Note that the test samples were similar to the controls when f was 0. In this case sSHM should not mark them as significantly different from NOC. A thousand samples were simulated for each value of f as follows to ensure that the canonical vector **a** was sparse:

$$\mathbf{x}_p = \mathbf{x}_h + f(e_i + e_j)\boldsymbol{\Sigma} \tag{6.6}$$

where $\mathbf{x}_h$ indicates a row vector that was drawn from $N(\mathbf{0}, \boldsymbol{\Sigma})$, i.e. a control sample. Note that half of the samples had a negative sign in f. The two abnormal variables were defined by

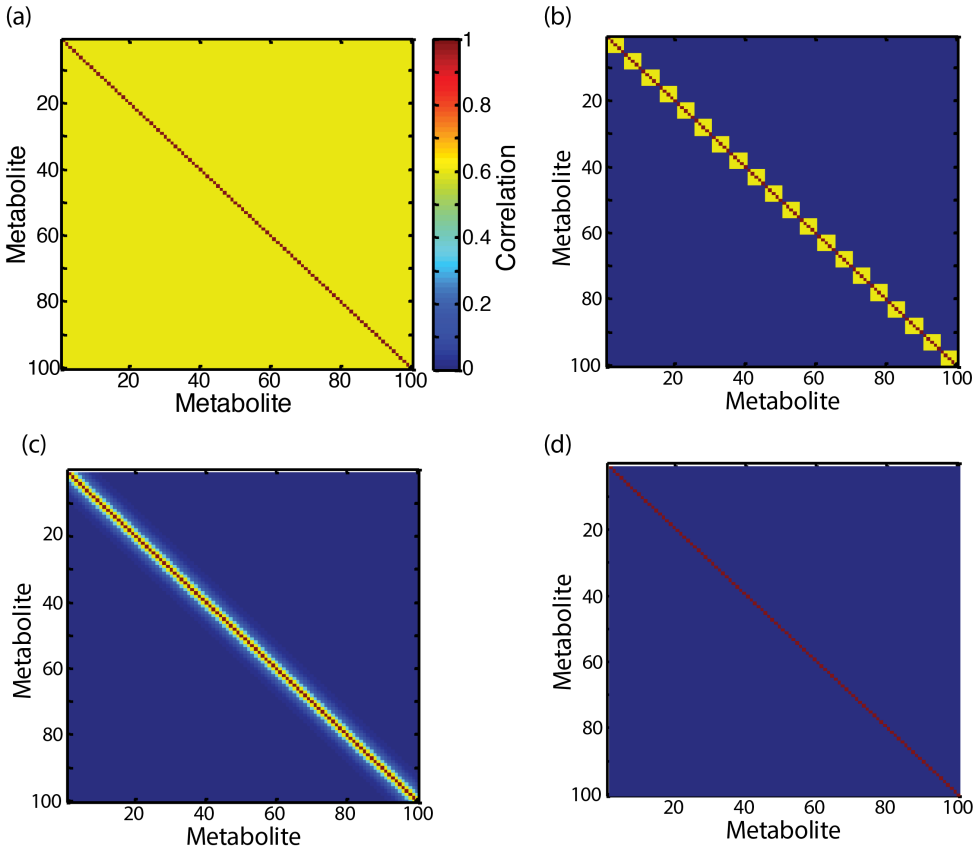

Figure 6.2 A graphical representation of **(a)** the common variance, **(b)** the block diagonal, and **(c)** the Toeplitz, correlation structures with $\rho = 0.6$ that were used in the simulation study. Structures **(a-c)** are equal to an independent structure when $\rho = 0$. This is shown in panel **d**.

unit vectors e_i and e_j, where i and j indicated their indices. For each simulated sample, the indices i and j were drawn from a uniform distribution. Prior to construction of the models, all simulated samples were autoscaled to the mean and standard deviation of their corresponding control samples.

6.3.2 Diagnosis of inborn errors of metabolism

To assess the value of sSHM for disease diagnosis, a set of urine samples of 193 healthy children and a set of 24 patients was measured using proton NMR spectroscopy. Eighteen patients were known to suffer from one of seven different IEM. For the other six patients, no IEM was diagnosed, but signals related to commonly prescribed drugs such as depakine and paracetamol were found in the NMR spectra. Note that a subject had to be between 4—12

years old to participate in the study and be of Dutch ancestry. An equal amount of males and females were selected. No other selection criteria such as lifestyle or diet were imposed. The study was approved by the medical ethical committee of the Radboud University Medical Centre in Nijmegen, The Netherlands. More details regarding the data can be found in section 5.3.1 [4]. This section also includes more information regarding the measurement and subsequent processing of the NMR data. Briefly, the regions 0.2-4.7 ppm and 5.0-10.0 ppm were selected for further analysis. Next, the urine NMR spectra were normalized to the creatinine concentration to correct for dilution effects. Equidistant binning with a bin size of 0.04 ppm was used to reduce the normalized data from 30888 measurements to 246 bins. Finally, Pareto scaling was applied to the data for reasons justified in section 6.4.3.

Since outliers in the set of controls can heavily influence an sSHM model, the spectra of the 193 healthy children were inspected using robust PCA. Seventeen samples with abnormal patterns related to dietary influences and drug intake were identified. These samples were marked as abnormal and used to validate the sSHM model since detection of abnormal patterns due to diet and drugs is in principle no different from the detection of abnormalities related to disease. The set of 24 patients was also used for this purpose. Additionally, another set of 56 samples from the remaining 176 healthy controls was used to validate the model. The remaining healthy control samples were used to train the sSHM model and define the NOC. Based on their value in equation 6.5, it was observed during cross-validation of the 120 training samples that 2 of these samples greatly differed from the others. Visual inspection of the data showed that this was due to bad baseline correction and water suppression. Therefore, these samples were excluded from further analysis.

As explained in section 6.2.2, the sSHM model improves the Mahalanobis distance in two ways. Firstly, a variable selection step is introduced. Secondly, the model is regularized by adding $\delta * \log(p)/n$ to the diagonal elements of covariance matrix $\mathbf{\Sigma}$. To clearly demonstrate the advantage of variable selection by sSHM compared to just the regularization of $\mathbf{\Sigma}$, the analysis was repeated 15 times where each time the data was concatenated with a block of 100 additional random variables (noise bins). The bins were normally distributed with zero mean and a standard deviation chosen such that the intensity of the noise bins was roughly equal to the median peak intensity of the control samples.

6.4 Results

6.4.1 Analysis of the simulated samples

As mentioned in section 6.3.1, the main properties of the sSHM model were studied by simulation. A large number of control and patient samples were simulated for different structures of the NOC. Below, the analysis of a single simulated sample is described to demonstrate the main properties of the method. Additional details of the simulation study such as the empirical type I and type II error rates of the sSHM model are presented in appendix material C.

Detection of abnormal samples

Here, we consider a single simulated sample ($f = 2$ in equation 6.6) for the case where the correlation matrix of the control samples had a Toeplitz structure with $\rho = 0.6$ and 50 training samples were available. The simulated sample was automatically compared to the NOC by sSHM. As shown in table 6.1, a significant difference could be observed due to variable selection. When using a significance level of 5%, for example, no significant difference was observed when all variables were taken into account: the effect of the abnormal variables was masked by the "noise" of the other variables. In contrast a clear difference was observed when only 2 variables were selected. This was expected since the simulated sample contained two abnormal variables. In real applications the number of variables that should be selected is unknown. As shown in the table, the simple standardization suggested in equation 6.5 ignored enough irrelevant variables such that a difference between the patient and the NOC was observed: the observed p-value was similar as the p-value found for the solution with 2 selected variables. Similar results were obtained for the other simulations (see appendix C).

Table 6.1 Accuracy of different sSHM tests for health monitoring of a simulated patient sample.

Number of selected variables		p-value
Manual	2	0.00
	10	0.00
	25	0.12
	75	0.12
	100	0.14
Automatic	1	0.00

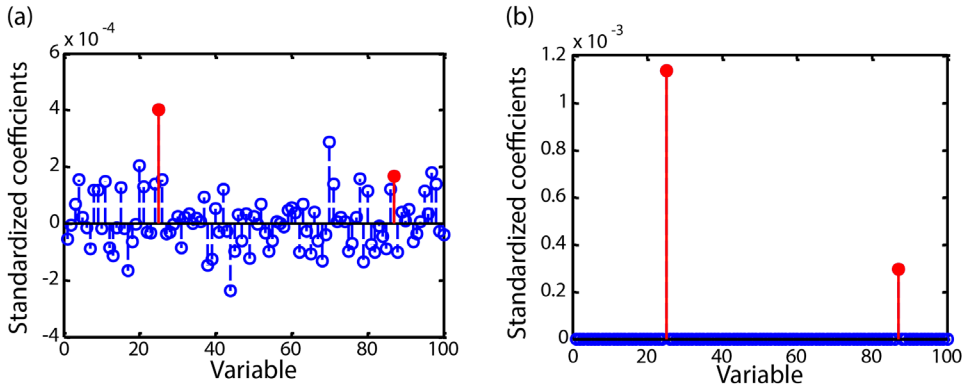

Figure 6.3 The canonical vectors of an sSHM model where **(a)** all variables were selected and **(b)** 2 variables were selected. These results were obtained for analysis of the same simulated patient sample as analysed in table 6.1. It was known that two variables were abnormal. These are marked red in the figure. The irrelevant (normal) variables are marked blue.

Identification of abnormal variables

The observed significant difference between the simulated sample and the NOC suggested that the sample should be further investigated to identify which variables were abnormal. Figure 6.3a shows the estimates of the coefficients of the canonical vector **a** that were obtained when all variables were selected. The abnormal variables (solid red circle, solid red line) could barely be distinguished from the irrelevant variables (blue empty circle, blue dotted line). Identification was greatly improved by the variable selection step in sSHM as shown in figure 6.3b. In this case the abnormal variables were easily identified since they were the only two variables that were selected.

In this simulation it was known that only 2 variables were abnormal. In real applications, however, this number is not exactly known and it can be difficult to select the right solution to inspect. Therefore, it is preferred to inspect multiple solutions together in the so-called solution path figure. The path found for this simulation is shown in figure 6.4. The solution path displays the estimated coefficients of **a** for all possible values of the ℓ_1-norm constraint (see equation 6.3). The lines in the figure are the paths of the coefficients shrinking towards zero as a function of the constraint. This allows a clinical practitioner to see how specific variables behaved as a function of the constraint. Note that the x-axis has been scaled to have a maximum value of one[3]. A value close to 0 means that the l_1-norm constraint was large and only a few variables were selected (had nonzero coefficient), while a value of 1

[3] The x-axis corresponds to the proportion of shrinkage ($|\mathbf{a}|_{\lambda_i}/|\mathbf{a}|_{\lambda_p}$) instead of the size of the l_1-norm constraint.

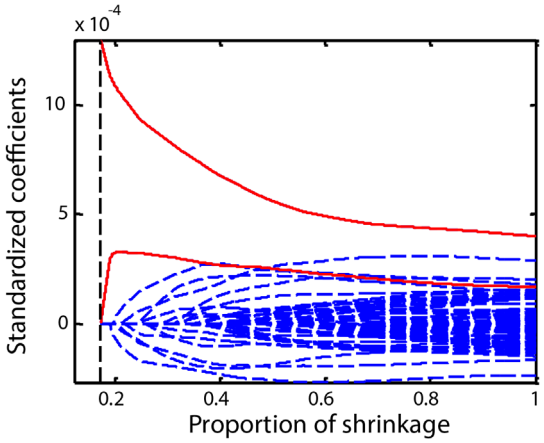

Figure 6.4 Solution path of an sSHM model applied to a simulated patient sample. The known abnormal variables are indicated by the solid red lines. The irrelevant (normal) variables are marked by dotted blue lines. The vertical dotted line indicates the solution found by automatic variable selection (equation 6.5).

indicates that all variables were selected. The solutions corresponding to figures 6.3a and 6.3b can be found at 0.19 and 1, respectively.

Moving from left to right in the solution path figure provides an intuitive way to see at which points a variable had a nonzero coefficient, i.e. was selected. In this case, the solution path clearly shows that the first two variables that were selected corresponded to the abnormal variables in the simulation (red solid lines). This demonstrates that the information in the solution path can be used to rank the variables, where it is assumed that the abnormal variables are among the first selected ones.

Note that the solution found by automatically variable selection with expression 6.5 selected only one of the two relevant variables. Therefore, inspection of the full solution path is preferred for identification of abnormal variables.

6.4.2 Comparison between SHM and sSHM

The main motivation for the development of sSHM was the unreliable identification of abnormal variables by SHM due to variable smearing (see appendix E). Below, the SHM and sSHM model are compared with respect to analysis of the simulated data. The analysis of the data by SHM was carried out according to the protocol described in section 5.3.3 [4] (see appendix E for more details regarding the SHM model). It was observed that SHM and sSHM had similar type I and type II error rates with respect to identification of significant

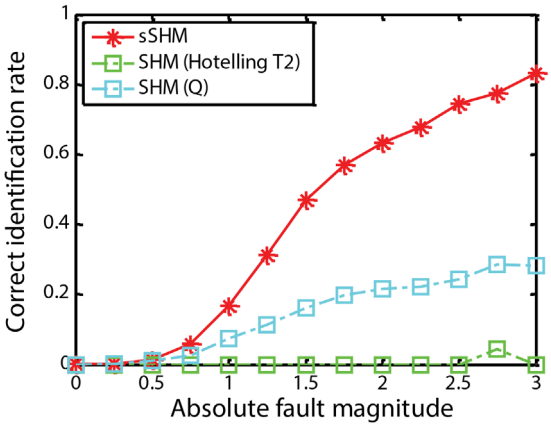

Figure 6.5 Correct identification of the abnormal variables in simulated patient samples by sSHM, and SHM as a function of the size of the abnormality.

differences between the simulated samples and the NOC. More details are provided in appendix E. Note that a big advantage of sSHM compared to SHM was that no parameter optimization was required: it is well known that the number of selected principal components can greatly influence the results of an SHM model [16].

As expected, sSHM greatly outperformed SHM with respect to identification of the abnormal variables in the samples. This is shown in figure 6.5 for the case that 50 training samples were available to define the NOC. To allow for a fair comparison between the different approaches it was assumed that all samples were correctly marked as significantly different from NOC. The x-axis in the figure expresses the difference between the NOC and the simulated patient as defined by constant f in equation 6.6. The y-axis shows the percentage of simulated samples in which the abnormal variables were correctly identified. Correct identification was achieved by sSHM when the first two selected variables were the abnormal ones. Identification of abnormal variables in SHM was achieved by studying the so-called contribution values [4]. Correct identification was achieved when the two abnormal variables had the highest contribution values. Note that SHM compared the patient to the NOC in two subspaces, namely a space spanned by the k selected principal components and the orthogonal or residual subspace. The green curve (labeled Hotelling T^2) in figure 6.5 corresponds to the contribution values found in the subspace spanned by the selected PCs and the cyan curve (labeled Q) corresponds to the values found in the residual subspace. From figure 6.5 it is clear that for sSHM and SHM the identification rate improved when the difference between the patient and the NOC was larger. However, as

shown in figure 6.5 the identification rate of SHM remained below 40% for the values of f considered. With sSHM considerable higher identification rates were achieved.

We also performed a comparison of SHM and sSHM to a third technique from the field of industrial process monitoring, namely sparse statistical process monitoring (sSPC; see appendix E). Similar to sSHM this method combines the MD with variable selection, although in a different way. It was found that sSHM outperformed sSPC in most scenarios.

6.4.3 Diagnosis of inborn errors of metabolism

Inborn errors of metabolism are a group of rare genetic defects that collectively occur in roughly 1 out of every 5000 individuals. Therefore, they are an important group of diseases to consider. Unfortunately, standard classification models such as (Partial Least Squares – Discriminant Analysis) cannot be used to diagnose these diseases since the number of training samples is extremely limited. In chapter 5, SHM was successfully used to analyze NMR data and diagnose several IEM [4]. Below, we will use the same data to show the value of sSHM in a practical example. As shown in table 6.2, the data did not only contain abnormal patient samples related to IEM, but also abnormalities related to diet and medication. These abnormalities had already been observed by visual inspection of the data and were also used to validate the sSHM approach since detection of abnormal patterns due to diet and medication is in principle no different from detection of metabolites related to a disease.

First, the data were analyzed after autoscaling. An improvement of roughly 10% in terms of the percentage of correctly identified patient samples (the specificity) was observed due to variable selection by sSHM. However, subsequent correct identification of the abnormal metabolites was hampered by the large influence of baseline signal on the model due to the autoscaling. Therefore, it was decided to apply Pareto scaling to the data instead. In this case, the abnormal metabolites were better identified (see below). As shown in table 6.3, however, the effect of variable selection on the number of correctly identified patient samples was less clear in this case. sSHM in combination with a large number of selected variable could also be used to identify the abnormal samples. This is attributed to the fact the the sSHM model also slightly regularizes the covariance matrix of the training samples $\mathbf{\Sigma}$ (see section 6.2.2). The original MD (equation 6.1) was not applicable to this data since the number of training samples was smaller than the number of variables. To clearly show the advantage of the variable selection step for the Pareto scaled data, the analysis was repeated 15 times. In each repetition the data was concatenated with a block of 100 irrelevant variables. As shown in table 3, the percentage of correctly classified samples was improved

Table 6.2 An overview of the abnormal samples that were investigated by NMR metabolomics in combination with sSHM. Note that the data was also analyzed in [4]. More details regarding the abnormalities can be found in this reference. The diagnostic metabolites that were used to diagnose each IEM are included in appendix F.

Disease (IEM)	n	Dietary	n	Medication	n	Other	n
3β-Hydroxy-Δ-C_{27}-steroid dehydrogenase deficiency	1	Cyclamate (artificial sweetener)	3	Depakine	1	Bacterial contamination	3
3-Methylcrotonyl CoA carboxylase deficiency	1	Fish	5	Paracetamol	5	High Taurine signal (cause unknown)	3
5-Oxoprolinuria	1			Piracetam	2		
Alkaptonuria	1			Sabril	1		
Cystinuria	2						
Formiminotransferase deficiency	2						
Isovaleric aciduria	1						

Table 6.3 Accuracy of different sSHM tests for health monitoring of 56 healthy and 41 abnormal samples for analysis. The original data as well as a concatenation of the data with 100 irrelevant bins was analysed. The sensitivity and specificity indicate the percentage of correctly classified healthy and abnormal samples, respectively. Note that high numbers of variables could be selected due to the regularization of covariance matrix $\boldsymbol{\Sigma}$ that was applied (see section 6.2.2). The reported results for the analysis with 100 irrelevant bins correspond to the average of 15 repetitions.

Number of irrelevant bins	Number of selected bins		Accuracy	Sensitivity	Specificity
0	Manual	2	95.2	100	90.5
		20	95.2	100	90.5
		50	95.2	100	88.1
		100	94.1	100	88.1
		200	94.1	100	90.5
	Automatic	1 - 246	95.2	100	90.5
100	Manual	2	95.2	100	90.5
		20	93.9	99.6	88.1
		50	93.0	98.8	87.1
		100	92.8	99.6	86.0
		200	92.5	99.9	85.0
	Automatic	1 - 346	94.6	100	89.1

by variable selection in this case. Note that the highest accuracy was obtained when the number of variables to select was automatically estimated by equation 6.5. All abnormal samples due to an IEM were correctly identified by sSHM. Four other samples were not

correctly identified. The abnormality in these samples was related to cyclamate (2x), paracetamol, and bacterial contamination.

After a sample had been marked as abnormal, the next step was to identify the abnormal metabolites. Again, it was observed that expression 6.5 could not be used in this step as indicated in e.g. figure 6.6a. Via inspection of the solution path figures, however, all IEM were successfully diagnosed and all dietary and medication abnormalities were successfully identified. Two examples are presented in figure 6.6. The solution paths of the other samples are shown in appendix F. In these figures the solid red lines indicate the variables (resonances in the NMR spectrum) that were known to be related to the IEM and other abnormalities in the data. These resonances should be selected first by the model. For clarity the relevant features are indicated in the NMR spectrum in panels b and d. In each case the variable selection rank (VS) indicates when the variable was selected by the sSHM model, where "1" corresponds to the first variable that was selected (the first nonzero coefficient in the left of panel a), etc. The IEM and other abnormalities were correctly diagnosed when enough of the relevant features had low VS values.

In the solution path of the first example in figure 6a the first four bins that were selected (VS 1 – 4) corresponded to resonances around 6.76 ppm and 3.64 ppm. This indicates that these resonances were abnormal compared to the NOC. The resonances indicate that a high amount of homogentisic acid was present in the urine of the patient, which is indicative for Alkaptonuria disease. In this case all known resonances (red lines) were selected first by the model and the diagnosis was clearly made. The second example is shown in figure 6c. This example is considered more difficult since it involved many resonances in many different parts of the NMR spectrum. The bins centred around 2.16, 5.14, 7.13, 7.31, 7.45, 9.67 and 9.78 ppm were clearly marked by the model as abnormal. All these features had VS values lower than 20. This suggested that the metabolites acetaminophen, acetaminophen-glucuronide, and acetaminophen-sulphate were present in high concentrations. This was caused by intake of paracetamol by this individual.

6.5 Discussion

In this work, sSHM was introduced for rapid screening of patients for many diseases. The properties of the method were investigated by simulation. Additionally, sSHM was successfully applied to real NMR metabolomics data to diagnose several IEM. The same data was previously investigated by the SHM model in chapter 5 [4]. Therefore, sSHM and SHM can be compared with respect to this application.

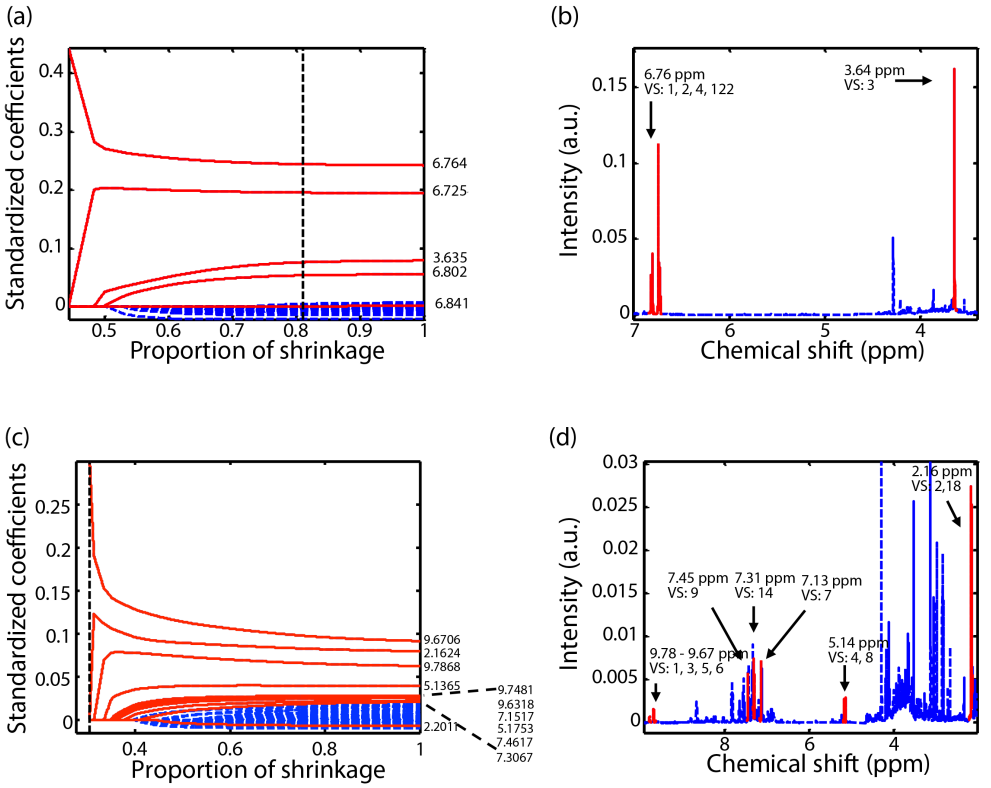

Figure 6.6 **(a, c)** Variable selection paths of sSHM models applied to the NMR data of a urine sample of two patients. The relevant diagnostic resonances are indicated by the solid red lines. The vertical dotted line indicates the solution of the maximum test (equation 6.5). **(b, d)** Plot of the NMR spectra highlighting the relevant diagnostic resonances and their variable selection rank (VS). The metabolites are markers for **(a, b)** Alkaptonuria disease, and **(c, d)** intake of paracetamol by the patient.

The methods had similar power to identify normal and abnormal samples. Compared to sSHM, SHM also correctly identified a significant difference between the four samples that were misclassified by sSHM. Next, SHM used contribution plots to identify the abnormal variables for the samples that were marked as significantly different. The contribution plots identified enough key variables as abnormal to make a diagnosis. However, interpretation of the plots was less straightforward because of two reasons: (1) the contribution values couldn't be reliably interpreted due to the so-called smearing effect (see appendix E), and (2) the contributions of the relevant variables differed by orders of magnitude making it difficult to visually inspect the results. In contrast, sSHM does not suffer from the smearing effect allowing for much better identification of the abnormal metabolites as shown in figure 6.6. This clearly shows that sSHM is a useful alternative to SHM in a clinical setting.

In this study it was shown that sSHM ranks the variables from abnormal to normal by variable selection. A clinical practitioner can use this information, which is visualized in the solution path figure, to diagnose a disease. Currently, however, no clear guidelines on how many of the top-ranked variables should be inspected are available. Equation 6.5 could clearly not be used for this purpose, nor is this expression based on firm statistical theory. It would be interesting to develop another strategy to obtain an upper bound. A possibly interesting approach in this respect is offered by stability selection [28]. Essentially, this method aggregates the results of multiple variable selection sSHM models that were applied to subsamples of the data. An attractive property of this method is that it allows for error control on the expected number of falsely selected variables. Additionally, for the LASSO model it has been shown that variable selection itself may improve markedly due to stability selection. This suggests that more abnormal variables might be among the first selected when sSHM is combined with stability selection.

The idea of combining the Mahalanobis distance with variable selection is not new. Several other methods have been proposed in the context of industrial process monitoring. These methods apply variable selection directly to the difference between the patient and the centre of the controls, i.e. $\mathbf{x}_p - \boldsymbol{\mu}_h$. However, this does not guarantee that the canonical variate $\mathbf{a} = \left[\mathbf{x}_\mathrm{p} - \boldsymbol{\mu}_h\right]\boldsymbol{\Sigma}^{-1}$ is sparse. The CV is directly penalized by sSHM. Because of this, sSHM is better able to identify abnormal variables in many situations. As mentioned in section 6.4.2 and appendix E this was indeed observed in our simulation study.

Similar to SHM, selection of the control samples is critical for practical application of sSHM for disease diagnosis. In principle, any metabolic variation that is not included in the NOC set will be marked as abnormal by the method. For example, in this study dietary abnormalities were not included in the NOC, and because of this some test samples were marked as abnormal based on their diet. If this is undesirable we suggest that a much larger cohort of control samples is used. If this cohort is randomly selected it should contain most common lifestyles and diets. Instead of using a larger cohort of samples, an interesting line of future research could also be to define the NOC at an individual level instead of a population level. It is expected that the resulting sSHM model would be able to detect more subtle abnormalities since no intra-individual differences are included in the NOC.

If the NOC is defined at a population level, it is important that the set does not contain outliers, such as people who are ill. Therefore, we propose to screen the control data by a robust PCA model before application of sSHM. Any NOC samples that are significantly different from the robustly defined NOC are excluded from further analysis. This approach was used to inspect the metabolomics data in section as described in section 6.3.2. An

alternative strategy to develop a robust method would be to replace $\boldsymbol{\mu}_h$ and $\boldsymbol{\Sigma}$ in equation 6.3 by robust estimators of location and scatter. However, this method is expected to perform similarly to the robust PCA approach.

6.6 Conclusion

In this work a new method – sSHM - for identification of disease biomarkers in individual patients was proposed. An advantage of this approach is that it can be used for screening of patients for a multitude of (rare) diseases. Additionally, the method offers perspectives in the framework of personalized health.

The sSHM method combines the Mahalanobis distance with variable selection. This way patient samples that are different from the expected healthy metabolic phenotypes (the normal operating conditions) can be better identified. Additionally, variable selection allows for identification of the abnormal variables when a patient is significantly different from NOC. As was shown by simulation and application to a real data set, sSHM can reliably identify patient samples and their individual disease biomarkes in many scenarios. The method has great advantages compared to SHM model that was proposed in chapter 5, especially with respect to identification of the individual disease biomarkes. Additionally, no optimization of the number of principal components is required as with SHM. However, the sSHM model is computationally more expensive.

sSHM is a general method that is not only applicable to metabolomics data. For example, most likely it could be of great value in industrial process control applications as well.

References

1. Li-Pook-Than, J. and M. Snyder, *iPOP goes the world: integrated Personalized Omics Profiling and the road towards improved health care.* Chemistry & biology, 2013. **20**(5): p. 660-666.
2. Chen, R., et al., *Personal Omics Profiling Reveals Dynamic Molecular and Medical Phenotypes.* Cell, 2012. **148**(6): p. 1293-1307.
3. Ellis, D.I., et al., *Metabolic fingerprinting as a diagnostic tool.* Pharmacogenomics, 2007. **8**(9): p. 1243-1266.
4. Engel, J., et al., *Towards the Disease Biomaker in an Individual Patient Using Statistical Health Monitoring.* PLoS ONE, 2014. **9**(4).
5. Engelke, U.F.H., et al., *Handbook of ^{1}H-NMR spectroscopy in Inborn Errors of Metabolism: Body fluid NMR spectroscopy and Brain MR spectroscopy*. 3 ed. 2014: Heilbronn: SPS Verlagsgesellschaft.
6. Nicholson, J.K., et al., *Metabolic phenotyping in clinical and surgical environments.* Nature, 2012. **491**(7424): p. 384-392.
7. Lindon, J.C., E. Holmes, and J.K. Nicholson, *Metabonomics in pharmaceutical R & D.* FEBS Journal, 2007. **274**(5): p. 1140-1151.
8. Madsen, R., T. Lundstedt, and J. Trygg, *Chemometrics in metabolomics—A review in human disease diagnosis.* Analytica Chimica Acta, 2010. **659**(1–2): p. 23-33.
9. Smolinska, A., et al., *Current breathomics—a review on data pre-processing techniques and machine learning in metabolomics breath analysis.* Journal of breath research, 2014. **8**(2): p. 027105.
10. Beckonert, O., et al., *Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts.* Nature protocols, 2007. **2**(11): p. 2692-2703.
11. Trygg, J., E. Holmes, and T. Lundstedt, *Chemometrics in Metabonomics.* Journal of Proteome Research, 2006. **6**(2): p. 469-479.
12. Johnson, J.M., et al., *A practical approach to detect unique metabolic patterns for personalized medicine.* The Analyst, 2010. **135**(11): p. 2864-2870.
13. Bersimis, S., S. Psarakis, and J. Panaretos, *Multivariate statistical process control charts: an overview.* Quality and Reliability Engineering International, 2007. **23**(5): p. 517-543.
14. Alcala, C.F. and S. Joe Qin, *Analysis and generalization of fault diagnosis methods for process monitoring.* Journal of Process Control, 2011. **21**(3): p. 322-330.
15. Van den Kerkhof, P., et al., *Analysis of smearing-out in contribution plot based fault isolation for Statistical Process Control.* Chemical Engineering Science, 2013. **104**(0): p. 285-293.
16. Ramaker, H.-J., et al., *The effect of the size of the training set and number of principal components on the false alarm rate in statistical process monitoring.* Chemometrics and Intelligent Laboratory Systems, 2004. **73**(2): p. 181-187.
17. De Maesschalck, R., D. Jouan-Rimbaud, and D.L. Massart, *The Mahalanobis distance.* Chemometrics and Intelligent Laboratory Systems, 2000. **50**(1): p. 1-18.
18. Mardia, K.V., J.T. Kent, and J.M. Bibby, *Multivariate analysis*. 1979: Academic Press.
19. Hastie, T.T., R.; Friedman, J., *The Elements of Statistical Learning*. 2001, New York, NY, USA: Springer New York Inc.
20. Wu, M.C., et al., *Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection.* Bioinformatics, 2009. **25**(9): p. 1145-1151.
21. Wu, M.C., *Statistical methods for high-dimensional genomic data*, in *Ph.D. Dissertation*2009, Harvard University.
22. Schäfer, J.S., K, *A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics.* Statistical Applications in Genetics and Molecular Biology, 2005. **4**(1).
23. Wang, K.B. and W. Jiang, *High-Dimensional Process Monitoring and Fault Isolation via Variable Selection.* Journal of Quality Technology, 2009. **41**(3): p. 247-258.
24. Zimek, A., E. Schubert, and H.-P. Kriegel, *A survey on unsupervised outlier detection in high-dimensional numerical data.* Stat. Anal. Data Min., 2012. **5**(5): p. 363-387.
25. Zou, C. and P. Qiu, *Multivariate Statistical Process Control Using LASSO.* Journal of the American Statistical Association, 2009. **104**(488): p. 1586-1596.
26. Capizzi, G. and G. Masarotto, *A Least Angle Regression Control Chart for Multidimensional Data.* Technometrics, 2011. **53**(3): p. 285-296.
27. Joe Qin, S., *Statistical process monitoring: basics and beyond.* Journal of Chemometrics, 2003. **17**(8-9): p. 480-502.

28. Meinshausen, N. and P. Bühlmann, *Stability selection.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2010. **72**(4): p. 417-473.
29. Wang, Q.J., *LH moments for statistical analysis of extreme events.* Water Resources Research, 1997. **33**(12): p. 2841-2848.
30. Hosking, J.R.M., *L-Moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics.* Journal of the Royal Statistical Society. Series B (Methodological), 1990. **52**(1): p. 105-124.
31. Mai, Q., H. Zou, and M. Yuan, *A direct approach to sparse discriminant analysis in ultra-high dimensions.* Biometrika, 2012.

Appendices

A. The effect of noise variables on the Mahalanobis distance

It can be shown that MD^2 is the optimal single-test statistic when a disease affects all variables simultaneously. However, often more structured abnormalities such as shifts in only some of the variables are expected, i.e. only a few metabolites are related to the disease. In this case all the other variables that were measured can be considered as irrelevant or noise. The MD^2 is not optimal for such shifts as shown using the simulation presented in figure A.1 [24]. The simulated data contained 100 variables and 1000 control samples drawn from a standard normal distribution. The patient samples were drawn from the sample distribution and subsequently the values of a subset of the variables were set to 2, i.e. they differed from the controls on these variables. The number of abnormal variables was varied from 1 to 100; 1000 patient samples were simulated in each case. As shown in the figure, the MD^2-statistic can hardly detect patients with small abnormalities in only a subset of the measurements. Once the number of abnormal measurements is above 20 (in this example), the patients become the highest scoring samples. Zimek *et al.* highlight that many other outlier detection routines suffer from the same problem [24]. This example shows that detection of significant differences between a sample and the NOC may be improved by variable selection: the accumulating noise in irrelevant variables is reduced by selecting most (all) abnormal variables and only a limited number of irrelevant variables.

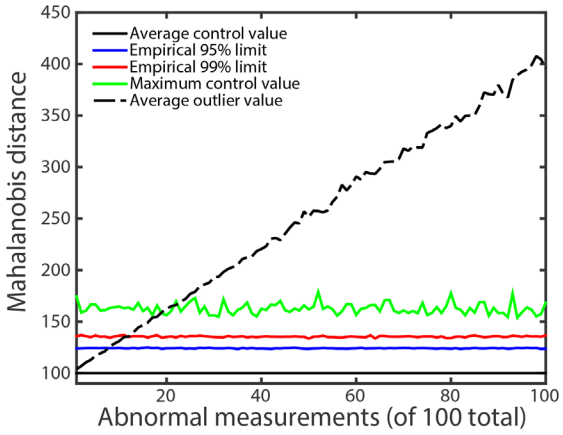

Figure A.1 Outlier detection based on the Mahalanobis distance in simulated data. Note that the MD couldn't identify outliers with only a few abnormal variables.

B. Significance testing with the sparse Mahalanobis distance

After selecting a specific number of variables a significance test is required to determine if the patient indeed differs from NOC. As mentioned above, an *F*-distribution is often used to determine the upper limit of MD^2 when all p variables are taken into account. However, this distribution is not applicable since we combine the MD-statistic with variable selection in this work. Even for healthy control samples there is a high probability that the sample has extreme (outlying) values in a subset of the variables. For example, for p independently normally distributed dimensions, the combined probability of an object appearing to be normal ($\alpha = 5\%$) in every single dimension is 0.95^p. For $p = 10$ and $p = 100$ the probability equals 59.9% and 0.592%, respectively. In other words, every object is extreme in at least one dimension in high-dimensional distributions. Because of this, we use extreme value theory to determine an upper limit for MD_i^2 and $MD_{\max}^2$.

Fitting the Generalized Extreme Value distribution.

The Fisher-Tippet theorem states that the distribution of extreme values can be described by the generalized extreme value (GEV) distribution [29]. The GEV has the following cumulative distribution function:

$$\begin{aligned} F(x) &= exp\left\{-\left[1-\frac{\kappa}{s}(x-\xi)\right]^{\frac{1}{\kappa}}\right\} \quad \kappa \neq 0 \\ F(x) &= exp\left\{-\exp\left[-\frac{1}{s}(x-\xi)\right]\right\} \quad \kappa = 0 \end{aligned} \tag{B.1}$$

or in inverse form

$$\begin{aligned} x(F) &= \xi + \frac{s}{\kappa}[1-(-\ln F)^{\kappa}] \quad \kappa \neq 0 \\ x(F) &= \xi - s\ln(-\ln F) \quad \kappa = 0 \end{aligned} \tag{B.2}$$

where κ is a parameter that controls the shape of the distribution, ξ is a location parameter and s is a scale parameter [29]. Note that x refers to the value for MD_i^2 or MD_{max}^2 that was found for a specific sample (patient or control). Note that equation B.2 can be used to determine the upper limit for the MD_i^2 or MD_{max}^2-statistics defined in equations 6.4 and 6.5 in the main text. Typically a value of $F = 0.95$ is chosen corresponding to a 5% significance level.

In order to use equations B.1 and B.2 the parameters κ, ξ, and s need to be estimated from the training data (a set of control samples). Typically a leave-one-out cross-validation

(LOO-CV) procedure is used to determine what values for MD_i^2 or MD_{max}^2 can be expected for samples that match the NOC. These values are ordered in ascending order. Maximum likelihood can be used to fit a GEV to these cross-validated values [29]. However, in cases when the sample size is small this method is known to be suboptimal. In this work we use the method of L-moments instead. Below, we present the required equations to be able to implement the method. More details can be found in numerous references such as [30].

The parameters of the GEV distribution are easily computed using L-moments via the following relationships:

$$\begin{aligned} \kappa &= 7.8590c + 2.9554c^2 \\ s &= \kappa l_2/[\Gamma(1+\kappa)(1-2^{-\kappa})] \\ \xi &= l_1 + \left(\frac{s}{\kappa}\right)[\Gamma(1+\kappa)-1] \end{aligned} \tag{B.3}$$

where $c = \frac{2l_2}{l_3+3l_2} - \frac{\ln(2)}{\ln(3)}$, and $\Gamma()$ is the Gamma function. The values l_1, l_2, and l_3 are the first three L-moments. The L-moments have a simple interpretation as measures of location, dispersion and shape of the data. The first three sample L-moments are given by:

$$\begin{aligned} l_1 &= b_0 \\ l_2 &= 2b_1 - b_0 \\ l_3 &= 6b_2 - 6b_1 + b_0 \end{aligned} \tag{B.4}$$

where b_r is the so-called *r*-th probability weighted moment defined by:

$$b_r = \frac{1}{r+1}\sum_{j=r+1}^{n} \binom{j-1}{r} x_j / \binom{n}{r+1} \tag{B.5}$$

Note that x_j corresponds to the *j*-th value in the sequence of ordered and cross-validated values of MD_i^2 or MD_{max}^2 of the training samples.

C. Investigation of the type I and type II errors of sSHM

A large number of control and patient samples were simulated for different structures of the NOC (see section 6.3.1). Either 50 or 1000 healthy control samples were used to define the NOC. The other samples were used to study the type I and type II error rates of the MD_i^2 or MD_{max}^2-statistics with respect to detection of significant differences between a sample and the NOC. It was observed that the type I and type II errors were similar across all

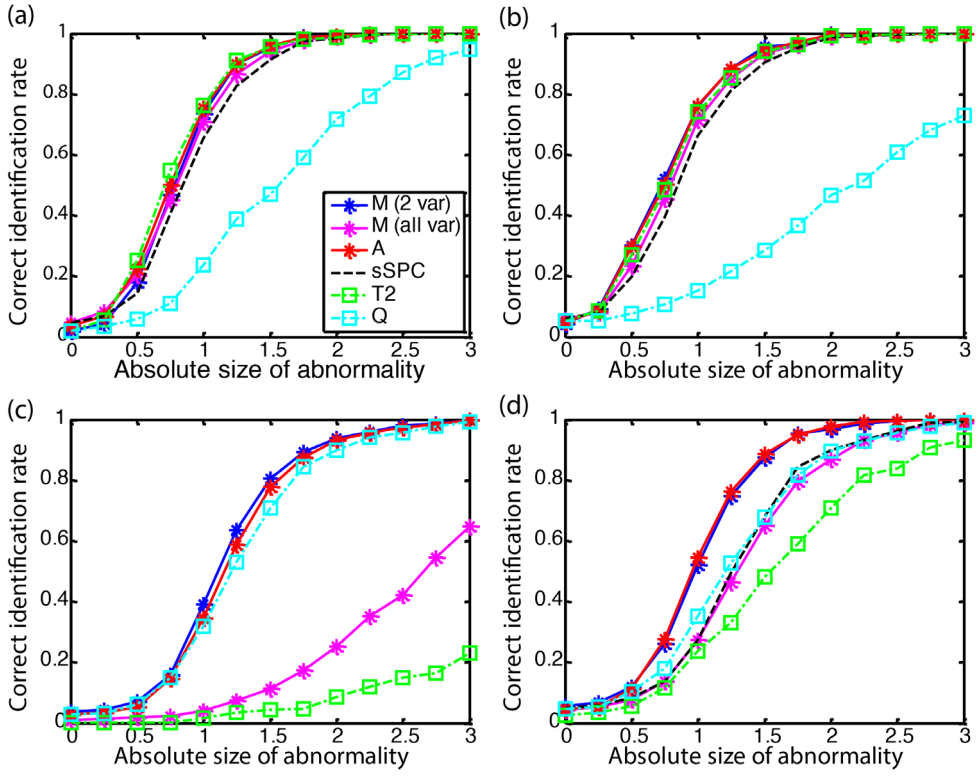

Figure C.1 The power of the sSHM, sSPC, and SHM (T2, Q) tests as a function of the size of the abnormality. The sSHM test was used with manual (M) and automatic (A) specification of the number of variables that should be selected. Panels **a – d** correspond to different configurations of the number of training samples and variables in the data; n = 50 and p = 10 in panel **a**; n = 1000 and p = 10 in panel **b**; n = 50 and p = 100 in panel **c**; and n = 1000 and p = 100 in panel **d**.

simulations with different correlation structures of the NOC. As an example we show the results that were obtained when the correlation structure corresponded to a Toeplitz matrix with $\rho = 0.6$.

Investigation of the empirical type I error

A type I error occurred when a significant difference between a healthy sample and the NOC was observed. A sample was marked as significantly different when the value of the test statistic (MD_i^2 or MD_{max}^2) exceeded the cut-off value defined by equation B.2.

The results are shown in figure C.1 for different combinations of number of training samples and variables. The x-axis in the figure corresponds to the parameter f defined in

equation 6.6. When $f = 0$ the simulated samples had the same characteristics as the controls, i.e. these samples were healthy and could be used to study the type I error. The correct identification rate displayed on the y-axis is defined as the percentage of simulations in which the method marked the simulated sample as significantly different. In other words, the y-value in the figure at $f = 0$ directly corresponds to the empirical type I error. Note that the curves labeled as sSPC, T^2, and Q in the figure were obtained with other methods than sSHM. These curves are discussed in appendix E where sSHM is compared to these other methods. Here, we focus on the results of the MD_i^2 and $MD_{\max}^2$ statistics.

From the figure it is clear that for all tests the type I error was reasonably close to the chosen significance limit of 5%. Similar patterns were observed for other significance levels, i.e. 1% and 10%. This indicates that the cut-off-point of the test could indeed be estimated from a GEV as in equation B.2. The GEV distribution, however, can be heavily skewed. Because of this, the fit became less accurate for smaller numbers of training samples. However, the estimated cutoff values were still more accurate compared to the values that were directly based on the empirical distribution of the test statistic (results not shown). Additionally, the resulting inaccuracies in the cut-off value were similar to what was observed for well-established upper limits of the statistics used in SHM. Therefore, the estimation of the cutoff-point by a GEV distribution as used here was deemed acceptable to control the type I error rate.

For regional flood frequency analysis the so-called method of LH-moments is often used to estimate high percentile quartiles of GEV distributions when the sample size is small. Sometimes this approach results in a more accurate estimator of the cutoff values compared to the method of L-moments we used. In these particular simulations this did not seem to be the case and therefore the method was not explored further (results not shown).

Investigation of the type II error

The results shown in figure C.1 for $f > 1$ were obtained for simulated samples that differed from NOC. Large values of f correspond to large differences as specified in equation 6.6. Because of this, the observed identification rate (power) for these samples could be used to study the empirical type II error rate, where the error rate is equal to 1-power. The lowest type II error was observed when the test used only 2 variables. This was expected since the patient samples were simulated such that only 2 variables were abnormal. Inclusion of irrelevant variables (i.e. when more variables were selected) had a negative effect on the type II error rate. For example, as shown in figure C.1, the error rate was considerably higher when all variables were selected. This effect was clearer when the data contained 100

variables instead of 10 and when the number of training samples was smaller. In practice the number of abnormal variables is unknown. This makes it challenging to select an appropriate number of variables such that a difference between the patient and the NOC can be best detected.

As proposed in equation 6.5 in the main text, this number can be estimated from the data. The result of the automatic selection approach is the curve labeled A in figure C.1. As shown in the figure, the type II error rate of this test was very close to what can be achieved when the number of important variables was known a-priori. This can be seen in figure C.1 by comparing the curve of the test with 2 selected variables to the curve of the automatic sSHM test. In other words, the simple standardization suggested in equation 6.5 is a simple way to select a suitable number of variables. It should be realized, however, that the maximum sSHM test was designed to negate the negative influence of irrelevant variables. It is not guaranteed that all of the relevant variables are selected.

D. Identification of abnormal variables

After a sample was marked as significantly different from NOC the next step was to identify which variables were abnormal. The simulation study was also used to determine how well sSHM could identify the abnormal variables under a wide range of different structures of the NOC. Again, similar results were observed for the different structures. As an example we show the results that were obtained when the correlation structure corresponded to a Toeplitz matrix with $\rho = 0.6$.

The results are shown in figure D.1. The x-axis in the figure expresses the difference between the NOC and the simulated patient as defined by constant f in equation 6.6. The y-axis shows the percentage of simulated samples in which the abnormal variables were correctly identified. Correct identification was achieved by sSHM when the first two selected variables were the abnormal ones (the simulation was constructed such that 2 variables were abnormal). Note that to generate the figure it was assumed that all samples were correctly marked as significantly different from NOC. From the figure it is clear that the correct identification rate improved for larger values of f. By comparing the top row of panels to the bottom row it can be seen that correct identification was not heavily influenced by the number of irrelevant variables in the data set.

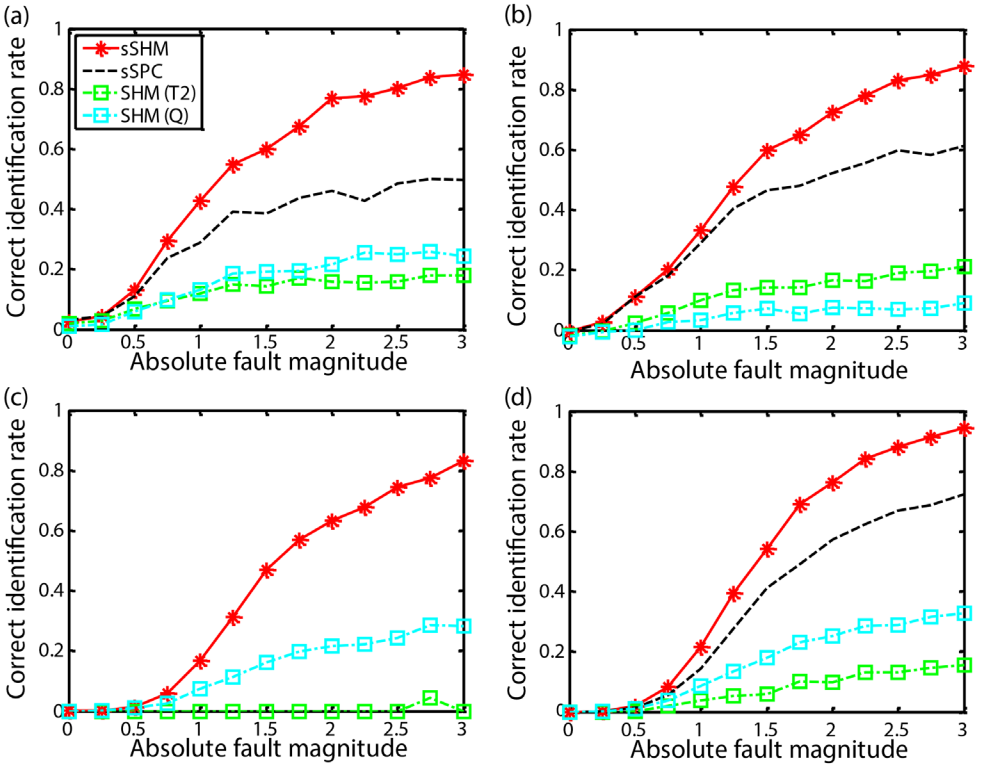

Figure D.1 Correct identification of the abnormal variables by sSHM, sSPC, and SHM as a function of the size of the abnormality. Panels **a– d** correspond to different configurations of the number of training samples and variables in the data; n = 50 and p = 10 in panel **a**; n = 1000 and p = 10 in panel **b**; n = 50 and p = 100 in panel **c**; and n = 1000 and p = 100 in panel **d**.

E. Comparison of sSHM to other methods

Comparison between sSHM and SHM

As mentioned in section 6.1, SHM can also be used to compare a patient sample to the NOC. In contrast to sSHM, the method combines the MD with dimension reduction instead of variable selection. More specifically, principal component analysis (PCA) is used to regularize the Mahalanobis distance and compare a sample to the NOC in two different subspaces [27]:

$$MD^2 = (\mathbf{x}_P)\boldsymbol{\Sigma}^{-1}(\mathbf{x}_p)^T = (\mathbf{x}_p)\mathbf{P}\boldsymbol{\Lambda}^{-1}\mathbf{P}^T(\mathbf{x}_p)^T + (\mathbf{x}_p)\widetilde{\mathbf{P}}\widetilde{\boldsymbol{\Lambda}}^{-1}\widetilde{\mathbf{P}}^T(\mathbf{x}_p)^T = \mathbf{T}^2 + \mathbf{T}_{RS}^2 \qquad \text{(E.1)}$$

where the columns in $\mathbf{P}$ and the diagonal elements in matrix $\mathbf{\Lambda}$ indicate the k eigenvectors (PC) and eigenvalues of $\mathbf{\Sigma}$ that are retained in the model, respectively. The matrices $\widetilde{\mathbf{P}}$ and $\widetilde{\mathbf{\Lambda}}$ indicate the $p-k$ residual eigenvectors and eigenvalues. The statistic T^2 is used to monitor the principal component subspace spanned by the k PCs. The residual space is typically monitored by the Q-statistic since $\mathbf{T}_{RS}^2$ cannot be applied when the control data has more variables than samples.

As shown in figure C.1 SHM and sSHM had similar type I and type II errors with respect to detection of significant differences between the NOC and a sample. However, sSHM greatly outperformed SHM with respect to identification of the abnormal variables. In SHM the values for T^2 and Q are decomposed into a sum of terms, each associated with one variable, called contributions to identify abnormal variables [14]. Unfortunately, the abnormal variables can increase the contribution of variables not influenced by the disease [14, 15]. This 'smearing' effect quickly leads to ambiguous diagnosis results. As shown in figure D.1 the correct identification rate of SHM was much poorer than that of sSHM due to the smearing effect. This shows that the sSHM model can be used more reliably to identify abnormal variables and diagnose as disease.

Comparison between sSHM and sSPC

The idea of combining the Mahalanobis distance with variable selection is not new. Several other methods have been proposed in the context of industrial process monitoring [25, 26]. Wang *et al.* use a forward-selection algorithm for this purpose [23]. The methods by Zou *et al.* and Capizzi *et al.* are closer to our approach. They also use an l_1-norm penalty to achieve variable selection. Since both approaches are very similar, we focus on the method developed by Zou *et al.* [25]. We will refer to this method as sparse statistical process control (sSPC). The defining feature of the sSPC approach is the following optimization problem:

$$\hat{\boldsymbol{\mu}} = argmin_{\mathbf{d}}(\mathbf{d}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{d}-\boldsymbol{\mu})^T, s.t. \sum_{i=1}^{p} |\mu_i|/|d_i| \leq \tau \qquad \text{(E.2)}$$

where $\mathbf{d} = \mathbf{x}_p - \boldsymbol{\mu}_h$. Problem E.2 is a generalized least squares problem that can be solved using various approaches such as least angle regression (LARS).

The goal of the above test is to determine whether the difference between the patient sample and the controls come from a distribution with $\boldsymbol{\mu} = \mathbf{0}$ (no difference), or a distribution with $\boldsymbol{\mu} \neq \mathbf{0}$ indicating that some variables have a different value compared to the average control

sample. Although, the direct penalization of $\boldsymbol{\mu}$ might seem like an intuitive sound choice, it does not guarantee that the canonical vector $\mathbf{a} = \boldsymbol{\mu}\boldsymbol{\Sigma}^{-1}$ is sparse. Note that sparse SHM aims to find a sparse estimate of the canonical vector directly. This should result in more accurate selection of the variables that can explain the difference between the patient and controls. Similar observations have been made when comparing differently penalized sparse LDA methods [31]. In other words, identification of abnormal measurements based on sSPC might be misleading.

As shown in figure C.1 the sSPC model had a slightly worse type II error rate with respect to identification of a significant difference between a sample and the NOC compared to SHM and sSHM. Additionally, the correct identification rate of the abnormal variables was slightly poorer as shown in figure D.1. Another big disadvantage of sSPC for analysis of –omics data is that the method is not applicable when the number of control samples is smaller than the number of variables. Therefore, no result was obtained when the number of training samples was 50 and the number of variables was 100 (panels c in figures C.1 and D.1).

F. Solution paths of the NMR data

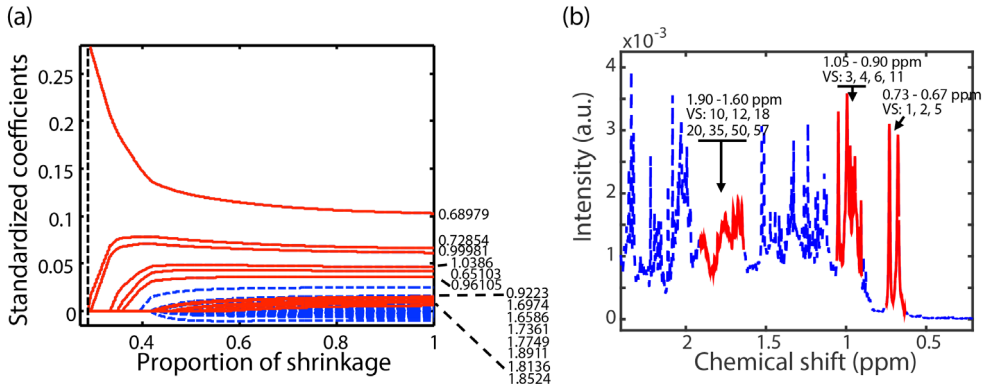

Figure F.1 **(a)** Variable selection path of an sSHM model applied to the **(b)** NMR data of a urine sample of a 3β-Hydroxy-Δ-C_{27}-steroid dehydrogenase deficiency patient. The solid red lines indicate the centre of the bins corresponding to resonances of the diagnostic metabolites dihydroxycholenic acid (0.67; 0.80 – 0.94; 1.60 – 1.90 ppm) and trihydroxycholenic acid (0.73; 0.80 – 0.94; 1.60 – 1.90 ppm). The dotted vertical line in panel **(a)** corresponds to the solution of the maximum test. The VS values in panel **(b)** correspond to the variable selection rank (VS) of the indicated resonances.

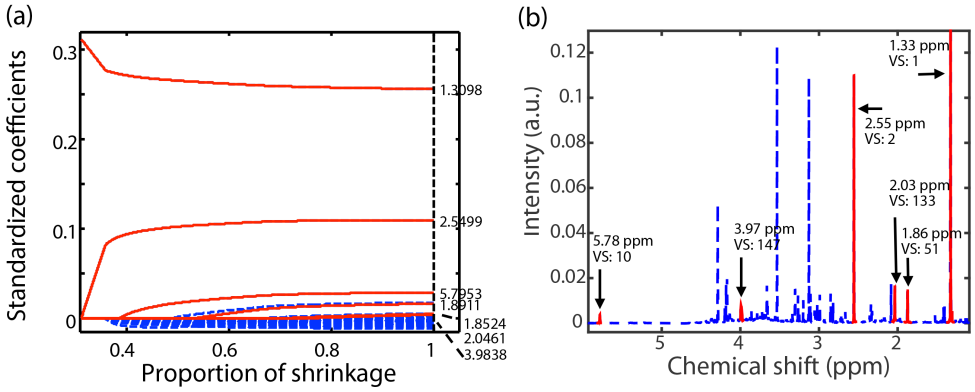

Figure F.2 **(a)** Variable selection path of an sSHM model applied to the **(b)** NMR data of a urine sample of a 3-Methylcrotonyl CoA carboxylase deficiency patient. The solid red lines indicate the centre of the bins corresponding to resonances of the diagnostic metabolites 3-methylcrotonylglycine (1.86, 2.03, 3.97, 5.78 ppm) and 3-hydroxyisovaleric acid (1.33; 2.55 ppm). The dotted vertical line in panel **(a)** corresponds to the solution of the maximum test. The VS values in panel **(b)** correspond to the variable selection rank (VS) of the indicated resonances.

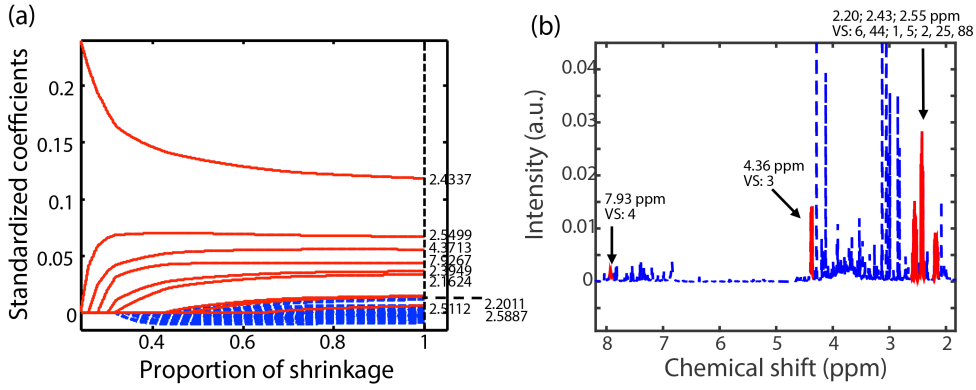

Figure F.3 **(a)** Variable selection path of an sSHM model applied to the **(b)** NMR data of a urine sample of a 5-Oxoprolinuria patient. The solid red lines indicate the centre of the bins corresponding to resonances of the diagnostic metabolite 5-oxoproline (2.20, 2.43, 2.55, 4.36, 7.93 ppm). The dotted vertical line in panel **(a)** corresponds to the solution of the maximum test. The VS values in panel **(b)** correspond to the variable selection rank (VS) of the indicated resonances.

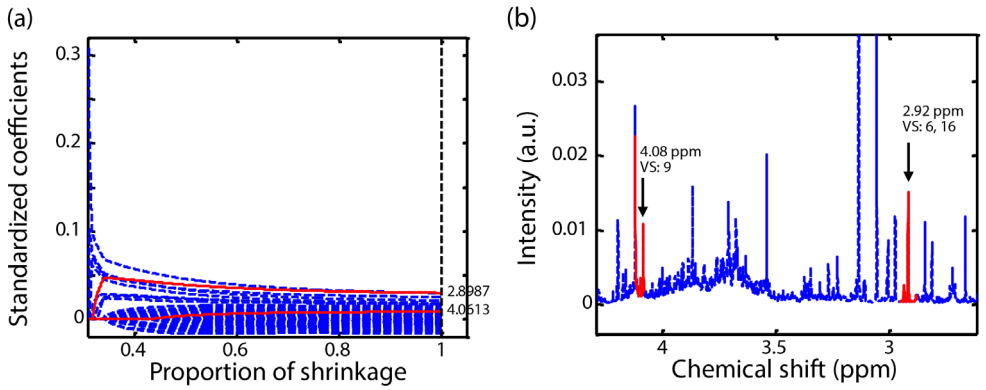

Figure F.4 **(a)** Variable selection path of an sSHM model applied to the **(b)** NMR data of a urine sample that is contaminated by bacteria. The solid red lines indicate the centre of the bins corresponding to resonances of the relevant metabolite *N*-methylhydantoin (2.92, 4.08 ppm). The dotted vertical line in panel **(a)** corresponds to the solution of the maximum test. The VS values in panel **(b)** correspond to the variable selection rank (VS) of the indicated resonances.

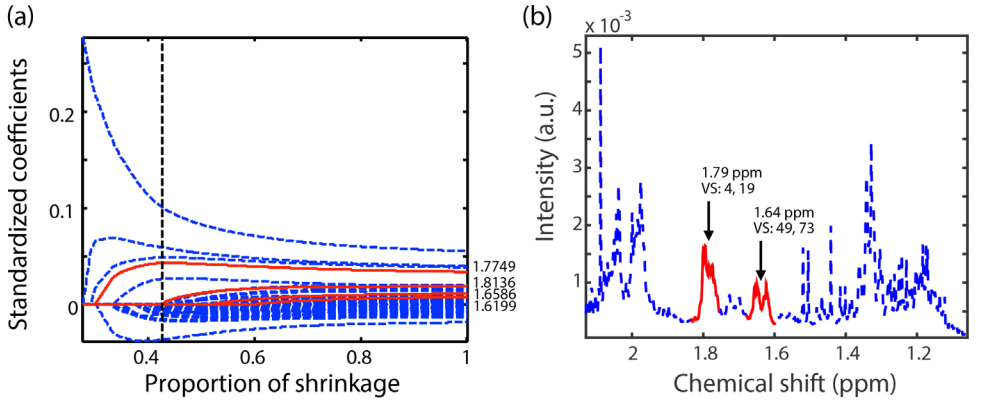

Figure F.5 **(a)** Variable selection path of an sSHM model applied to the **(b)** NMR data of a urine sample of a healthy individual who consumed an artificial sweetener. The solid red lines indicate the centre of the bins corresponding to resonances of the relevant metabolite cyclamate (1.64; 1.79 ppm). The dotted vertical line in panel **(a)** corresponds to the solution of the maximum test. The VS values in panel **(b)** correspond to the variable selection rank (VS) of the indicated resonances. Note that the bins that were selected first corresponded to unknown compounds with multiplets at 1.34 (VS 1, 2) and/or 1.99 (VS 1) ppm. Inspection of the NMR data by a clinical expert indeed confirmed that these resonances were abnormal with respect to the control samples.

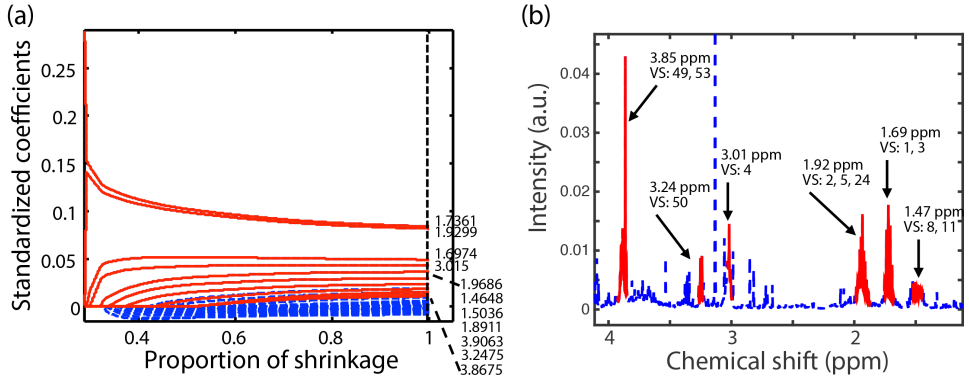

Figure F.6 **(a)** Variable selection path of an sSHM model applied to the **(b)** NMR data of a urine sample of a Cystinuria patient. The solid red lines indicate the centre of the bins corresponding to resonances of the relevant metabolites arginine (1.69, 1.92, 3.24, 3.85 ppm) and lysine (1.47, 1.72, 1.92, 3.01, 3.85 ppm). The dotted vertical line in panel **(a)** corresponds to the solution of the maximum test. The VS values in panel **(b)** correspond to the variable selection rank (VS) of the indicated resonances.

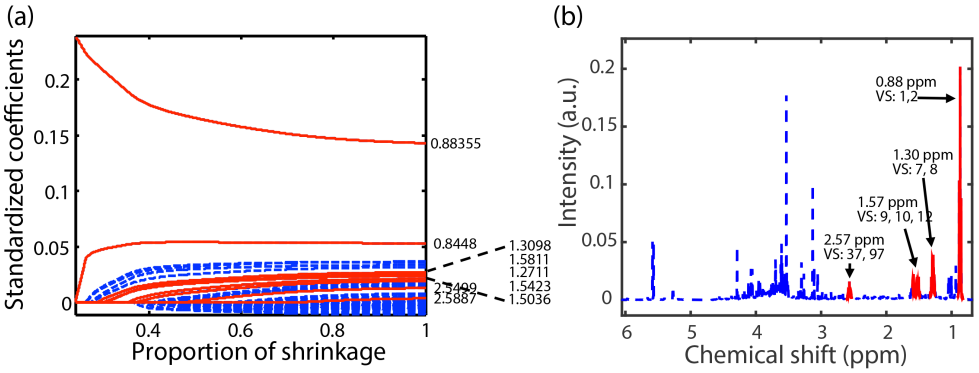

Figure F.7 **(a)** Variable selection path of an sSHM model applied to the **(b)** NMR data of a urine sample of a patient who was on depakine medication. The solid red lines indicate the centre of the bins corresponding to resonances of the relevant metabolite valproic acid (0.88, 1.30, 1.50, 2.44 ppm). The VS values in panel **(b)** correspond to the variable selection rank (VS) of the indicated resonances. Note that one variable was selected by the maximum test. The strong signal inbetween 5 and 6 ppm was selected with VS 4 and 5 and might be related to 4-Amino-5-hexenoic acid. This might be due to sabril intake by the patient.

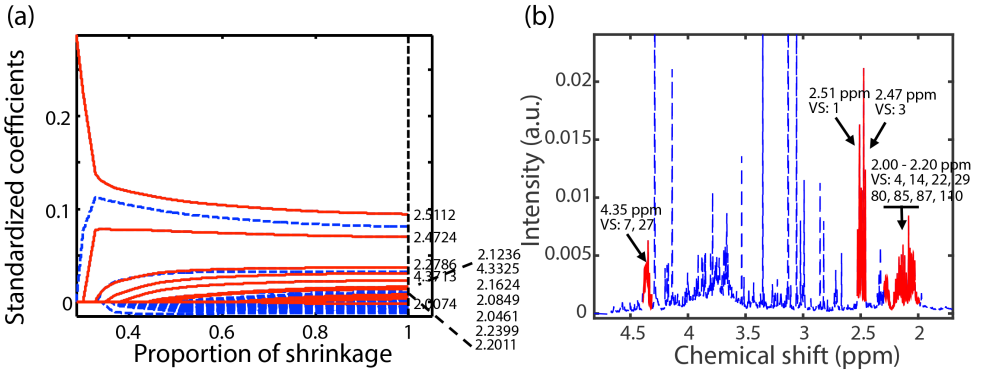

Figure F.8 **(a)** Variable selection path of an sSHM model applied to the **(b)** NMR data of a urine sample of a Formiminotransferase deficiency patient. The solid red lines indicate the centre of the bins corresponding to resonances of the diagnostic metabolites formiminoglutamic acid (2.00 – 2.22; 2.47 ppm) and hydantoin-5-propionic acid (2.00 – 2.22; 2.51 ppm). The bin corresponding to the multiplet around 7.80 – 7.92 ppm was the second bin (VS 2) selected by the model and is related to one of the diagnostic metabolites (model compound not available). The dotted vertical line in panel **(a)** corresponds to the solution of the maximum test.

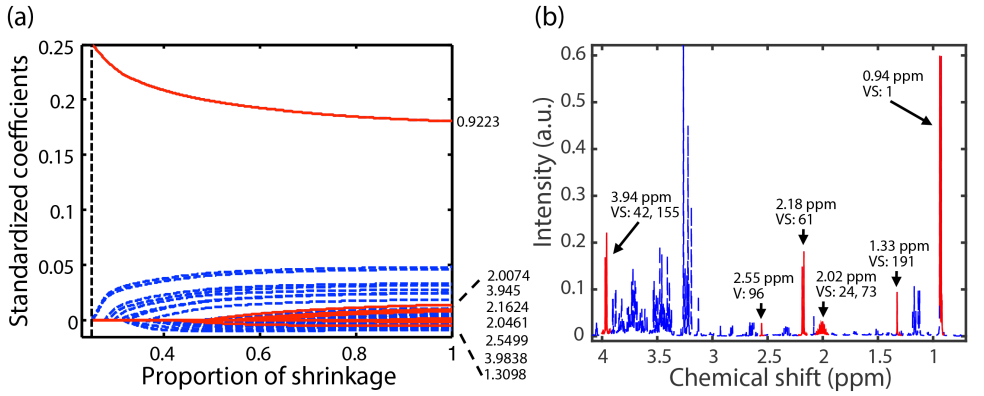

Figure F.9 **(a)** Variable selection path of an sSHM model applied to the **(b)** NMR data of a urine sample of a Isovaleric aciduria patient. The solid red lines indicate the centre of the bins corresponding to resonances of the diagnostic metabolites 3-hydroxyisovaleric acid (1.33, 2.55ppm) and isovalerylglycine (0.94, 2.02, 2.18, 3.94 ppm). The dotted vertical line in panel **(a)** corresponds to the solution of the maximum test. Large quantities of choline, acetylcholine and glucose were observed in the spectrum of the patient. Resonances due to these metabolites were first selected by sSHM explaining why many of the resonances related to 3-hydroxyisovaleric acid and isovalerylglycine had high VS values.

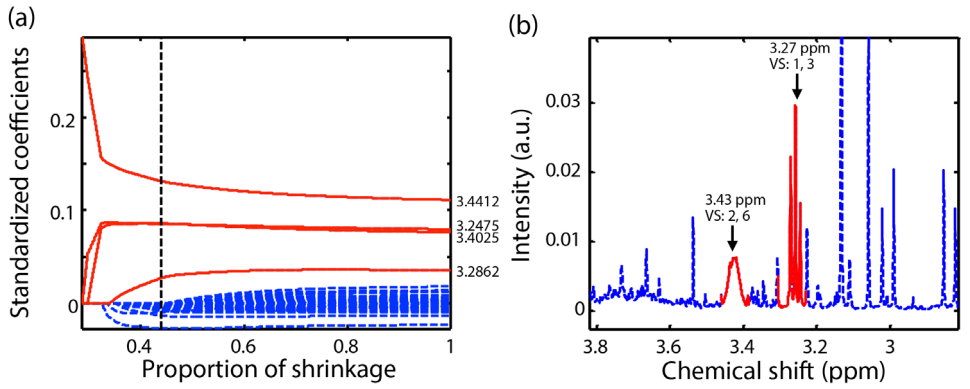

Figure F.10 **(a)** Variable selection path of an sSHM model applied to the **(b)** NMR data of a urine sample of a patient with high amounts of taurine in his urine (cause unknown).. The solid red lines indicate the centre of the bins corresponding to resonances of the relevant metabolite taurine (3.27, 3.43 ppm). The VS values in panel **(b)** correspond to the variable selection rank (VS) of the indicated resonances.

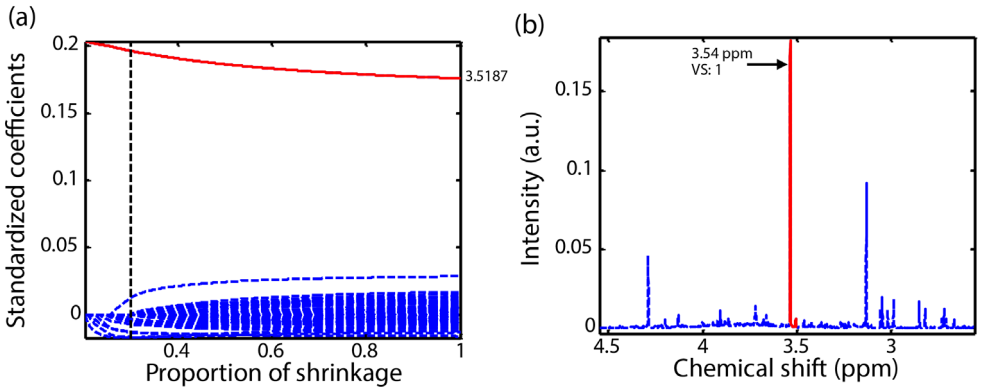

Figure F.11 **(a)** Variable selection path of an sSHM model applied to the **(b)** NMR data of a urine sample of an individual who just consumed fish.. The solid red lines indicate the centre of the bins corresponding to resonances of the relevant metabolite trimethylamine-oxide (3.54 ppm). The VS values in panel **(b)** correspond to the variable selection rank (VS) of the indicated resonances..

7

SUMMARY AND FUTURE PERSPECTIVES

Chemometrics on its way towards personalized health care

Metabolomics has great potential to contribute to a transformation of health care towards approaches that are tailor-made for specific individuals instead of being optimal for the average patient with a specific disease. This is underpinned by our observation that the normal value for a body fluid biomarker in an individual varies within narrow limits well within the reference range that clinical chemistry laboratories use to define normality for that parameter. This strongly suggests that defining individual reference ranges, thus defining a person's normal operating conditions, holds a great promise. It will allow earlier detection of deviations from the healthy state and diagnostics in an earlier phase of disease, but also more precise estimates of the patient's health state when he/she is recovering from a diseased state. That would be an important step for precision or personalized health care.

Data analysis is a current bottleneck that prevents such applications. Therefore, novel data analysis approaches were proposed in the previous chapters of this thesis. The main findings of this research are summarized here. Subsequently, it is shown that extension of the developed methodology for analysis of longitudinal data would be extremely interesting. Such data allows for assessment of individual responses to a biological perturbation by statistical modeling, i.e. truly personalized modeling.

7.1 Summary

In current clinical practice, most diagnoses and treatment plans have been designed for the average patient. Because of this, treatment can be very successful for some individuals, but not for others. The rapid development of high-dimensional analytical technologies is expected to overcome these limitations. These techniques allow for examination of a patient in unprecedented detail at a molecular level. It is expected that this will result in significant advances in health care at two levels: (1) a better understanding of biological processes whose mechanisms are currently unclear may be obtained, and (2) with improved understanding of these processes so-called personalized health care can be applied where diagnosis and treatment are completely catered towards the unique molecular characteristics of an individual patient.

As described in the general introduction provided in **chapter 1** *the goal of this thesis was to address several challenges in analysis of data generated by high-dimensional technologies for personalized health care*. Work mainly focused on analysis of untargeted metabolomics data. Metabolomics refers to the use of high-dimensional techniques for the global and unbiased measurement of metabolites in a sample such as urine or blood. This approach is expected to become extremely important for personalized health care because the concentrations of these metabolites provide detailed information regarding the patient's phenotype. However, currently our ability to carry out metabolomics experiments greatly outpaces our ability to analyze the acquired data. Development of novel data analysis approaches is essential to unlock the potential of metabolomics for personalized health care.

Pre-processing is a crucial first step in data analysis to remove unwanted variation such as baseline drifts, peak misalignments, etc. These artifacts can completely mask the relevant biological information in the data. Based on analysis of a Fourier transform-infrared benchmark data set it was shown in **chapter 2** how difficult it can be to select an appropriate pre-processing strategy for a specific data set. Only a small amount of the seemingly viable approaches improved subsequent statistical modeling, while many methods actually had a negative impact. Additionally, it was shown that all three widely used approaches for pre-processing selection have serious drawbacks; they may be time-consuming beyond practicability or may provide misleading results. It is clear that more efficient pre-processing selection approaches need to be developed.

Statistical modeling, e.g. to compare groups of samples to each other or diagnose a disease, is the second data analysis step. As described in **chapter 3** metabolomics data poses significant challenges for statistical modeling. Many traditional techniques are not

applicable, or their results are unreliable. One way to circumvent these issues is to impose some structure to the model with the aim to reduce its variance at the cost of some bias. This is known as regularization. Chapter 3 reviewed modern regularization methods that might be of use to the field of metabolomics. Specific attention was paid to regularized estimators of the covariance matrix since this matrix is such an essential element of many statistical techniques. Subsequently, the use of l_1 -norm constraints to combine model estimation and variable selection was described. The regularization approaches that were reviewed in this chapter are widely applicable and formed the basis for the statistical models that were developed in chapters 4 - 6.

A first example of regularization was considered in the context of analysis of metabolomics experiments that were measured according to an experimental design. Such experiments are important for personalized health care to investigate the influence of different factors and interactions on metabolism and gain more mechanistic understanding. Multivariate analysis of variance (MANOVA) is a traditional method for analysis of data with an underlying design. However, it cannot be applied to typical metabolomics experiments in which a large number of variables are measured on a few samples. ANOVA simultaneous component analysis (ASCA) is a recently proposed alternative to MANOVA for analysis in these cases. It was shown in **chapter 4** that the ASCA model implicitly assumes that metabolite concentrations are independent (uncorrelated) and have the same variance. This unrealistic assumption reduces the power of the method and hampers interpretation. Therefore, an improved model was proposed that is essentially a weighted average of the ASCA and ANOVA model. This was achieved by combining MANOVA with a modern regularized estimator of the covariance matrix. By means of simulation it was demonstrated that the regularized MANOVA (rMANOVA) model has higher power compared to ASCA under a wide range of alternative hypotheses. This was also observed when techniques were applied to urine ^{1}H-NMR metabolomics data originating from a nutritional intervention study.

Metabolomics studies used to compare groups of samples to each other (e.g. patient with a specific disease and controls). However, as argued in chapters 1, 5 and 6 this might be impractical in a clinical setting and a paradigm shift moving beyond such population-based approaches is required. In **chapter 5** it was shown that ideas from industrial process monitoring are extremely useful in this respect. Therefore, statistical health monitoring (SHM) was introduced, which adapts data analysis ideas from industrial process monitoring to the clinical setting. The SHM model can be used to automatically identify the disease response in the individual patient. This is achieved by comparing the measured data of an individual patient $(n = 1)$ to that of a group of controls. The model indicates metabolic differences between the patient and the controls. A tremendous advantage of

SHM is that only data of the control group is required to construct the model. Because of this, SHM is not disease specific and can e.g. be used for diagnosis of rare diseases. In chapter 5 it was shown that several inborn errors of metabolism (IEM) could be diagnosed by ^{1}H-NMR metabolomics in combination with the SHM model. Diagnosis of these rare diseases by conventional population-based approaches is challenging because typically not enough samples are available to for a useful comparison between the patient and controls. Additionally, a large number of diseases need to be considered.

The SHM model described in chapter 5 used the Mahalanobis distance (MD) in combination with regularization via principal component analysis (PCA) to compare the patient to the controls. Because of the PCA step, the model is applicable when the number of control samples is smaller than the number of variables. A major drawback is, however, that it allows for interactions between abnormal variables and normal variables. This is known in industrial process monitoring as the smearing effect. Because of this, reliable identification of the metabolic differences between patient and controls is challenging. Therefore, a novel SHM model was proposed in **chapter 6**. The key idea consists of regularizing the MD estimator by variable selection instead of dimension reduction. Variable selection was achieved by inclusion of an ℓ_1-norm constraint during estimation of the MD. By means of simulation it was shown that the resulting sparse SHM model could better identify the differences between patient and controls. The model was also successfully applied to ^{1}H-NMR metabolomics data for diagnosis of several IEM.

In conclusion, high-dimensional technologies pose many problems for data analysis. The main contributions of this thesis consist of a careful evaluation of current data pre-processing strategies and the development of three novel statistical models for analysis of metabolomics data for personalized health care. The new models were successfully applied to urine ^{1}H-NMR metabolomics data to assess the metabolic response of patients suffering from metabolic syndrome to four diets, and to inspect the health status of individual patients and diagnose several rare diseases. These results show that the developed methodology holds great promise in personalized health care. Follow-up studies are, however, required to further validate these results and to study the value of the developed methodology in other applications.

7.2 Future perspectives

Due the emergence of increasingly large data sets in science we expect that the methodology proposed in this thesis is widely applicable. Within the context of personalized health care

the application of the developed methodology for longitudinal studies seems especially interesting. Such studies are of interest because often it is not possible to predict a-priori which time point best captures a specific biological effect and at least several time-points need to be captured. Additionally, time is an important factor to understand biological processes. Finally, data with a longitudinal design potentially allows for separate assessment of individual responses to a biological perturbation by statistical modeling. In other words, models that are truly personalized can be applied.

7.2.1 Data pre-processing

Data pre-processing is an important aspect of any metabolomics experiment, including experiments with a longitudinal design. As mentioned above, selection of an appropriate pre-processing strategy is not an easy task and guidelines need to be developed. When the choice of a method cannot be driven by analytical and biological knowledge (e.g. centering and scaling the data to a specific time point), the use of data quality parameters to judge the quality of the pre-processed data seems a useful tool to select the optimal methods. Some of these parameters were evaluated in chapter 2. However, it is not expected that a single parameter can capture all aspects of data quality. Therefore, it should be investigated which parameters have to be combined and in what way this has to be done. The sum of ranking differences (SRD) method seems interesting in this respect [1, 2]. The method can be used for data with repeated observations (longitudinal data) [3, 4]. SRD was developed to rank different models (pre-processing approaches) based on multiple merits (data quality parameters) and automatically select the best model or collection of models if so desired. Because SRD combines multiple merits in a natural way, decisions on how to combine and weight the different merits are avoided. SRD was recently used in this way to find the optimal shrinkage parameter in a regression model [4]. Ideally, the use of data quality parameters for pre-processing selection should be integrated with current practice in metabolomics such as the use of quality control (QC) samples that are measured throughout the experiment [5]. For example, quality measures can be derived from the QC samples such as the % relative standard deviation of each metabolite feature. Hopefully, application of this approach to pre-processing will reveal trends regarding suitable and robust approaches for specific data types and applications. Based on this, more specific guidelines can hopefully be developed to improve consistency between different studies, different laboratories, etc.

7.2.2 Statistical modeling: biomarker discovery.

Many metabolomics experiments to detect biomarkers have a longitudinal design. This way processes such as disease resilience and disease progression can be studied. Also,

longitudinal data allows for assessment of individual differences, e.g. is the temporal metabolic response to a perturbation the same for all individuals involved in the study? Such information is essential to further improve our understanding of health and disease and application of personalized health care.

Currently, the number of methods to analyze data with a longitudinal design is very limited and often the dynamic nature of the experiment is not taken into account [6]. The rMANOVA model developed in chapter 4 seems a useful technique to study the effect of time, the perturbation, and their interaction. This model, however, is only applicable to cross-sectional data. By definition analysis of such data only provides information regarding the average response to a perturbation and no individual differences can be assessed. Individual differences are captured by longitudinal studies, but such data cannot be analyzed directly by rMANOVA since (1) the intra-individual correlations due to time are not taken into account, and (2) no individual time curves are fitted [7, 8]. In univariate statistics these problems are resolved by including so-called random effects to the model [7, 8]. This way, individual trajectories are incorporated in the model in a formal way such that questions about individual behavior may be considered. However, application of random effects for high dimensional data introduces challenges regarding the complex correlation structures between the repeatedly measured samples. It is not directly evident how the relationship between time and all the metabolites should be specified in such a model and most likely some form of regularization is required. Recently, Schelldorfer et al showed that regularization by an l_1-norm penalty was very useful in this respect [9]. They considered the case of a high dimensional design matrix. An et al have studied the case of high-dimensional response (the data matrix) and regularized the model by means of a factor analysis approach [10]. These papers might offer a good starting point for development of such approaches in metabolomics. Another interesting line of research would be to reduce the number of variables in the data by grouping metabolites that share common biological function (e.g. are related to each other via the same metabolic pathway) Subsequently traditional multivariate methods such as a mixed model could be applied. This strategy is similar to gene set enrichment and some initial approaches have been proposed in metabolomics [11, 12]. Currently, however, application of enrichment techniques is extremely challenging since a large number of metabolites in body fluids still remain unknown. However, this bottleneck has been recognized and much effort is undertaken to improve metabolite identification [13]. Therefore, it is expected that metabolite enrichment strategies will become more feasible in the future.

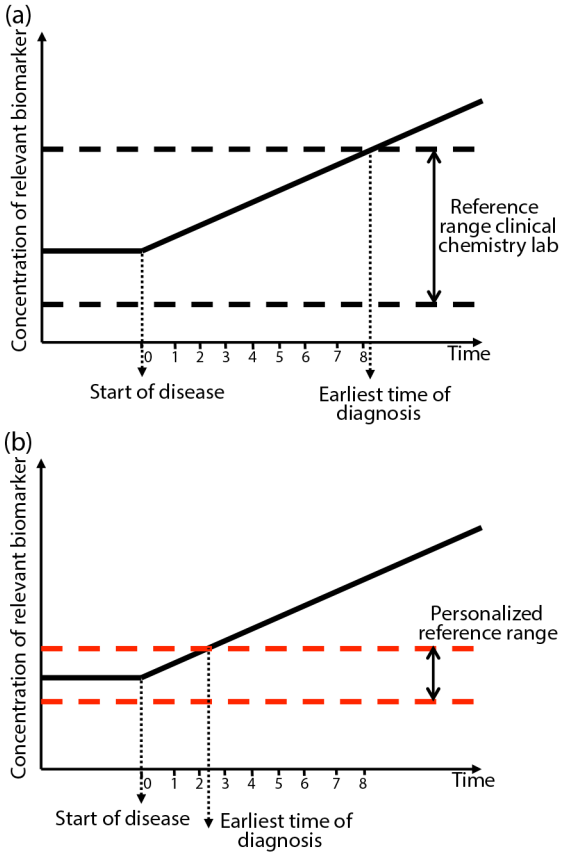

Figure 7.1 The concept of health monitoring using **(a)** population-based reference values for the relevant biomarker, and **(b)** personalized reference values.

7.2.3 Statistical modeling: personalized health monitoring

As shown in figure 7.1, longitudinal experiments also offer interesting possibilities for health monitoring by SHM. Firstly, by monitoring the physiological status of an individual through time the emergence of a disease may be detected before symptoms occur allowing for early and efficient treatment (see figure 7.1a, and application 3 in section 1.1.2) [14-16]. Secondly, longitudinal data may offer the possibility to compare abnormal/disease states to healthy states of the same individual instead of population estimates as in chapters 5 and 6 (figure 7.1b). Note that the expected healthy states or normal reference values were referred to as NOC (normal operating conditions) in these chapters. Based on the results of

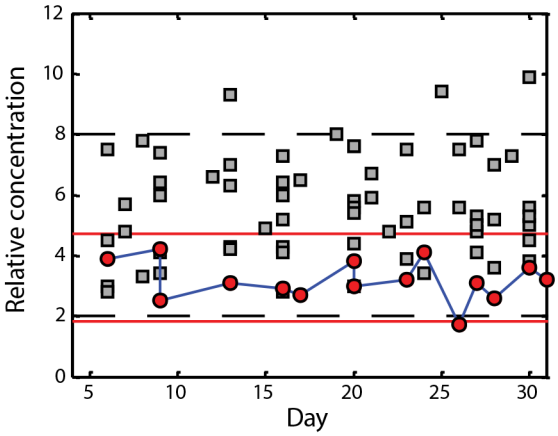

Figure 7.2 Relative concentration of pyruvic acid in urine (µmol/mmol creatinine) as observed in 5 individuals at 16 time points in January 2014. It is clear that the clinical reference values indicated by the black dotted lines match the observed metabolite concentrations for the population (grey squares and red dots), but are far from optimal for the individual marked in red. The red lines indicate more optimal personal reference values (NOC) for this specific individual. Such personal reference values will allow earlier and more accurate diagnostics.

a small pilot study[4] we expect that estimation of individual NOC's will allow for detection of subtle temporal abnormalities that would normally be hidden in the noise of the inter-individual variation. This is demonstrated in figure 7.2 for univariate analysis of the data from the pilot study, and in figure 7.3 by multivariate SHM analysis of the same data. More research is required to thoroughly test this hypothesis. Currently, several longitudinal studies that monitor the health status of individuals are under way, for example by the Stanford center for genomics and personal medicine. It would be extremely interesting to analyze the acquired data by SHM to compare individual and population estimates of the NOC.

[4] In the pilot study 5 healthy individuals were followed for 1 month (January 2014). The individuals all lived in the Nijmegen area and were matched in terms of age, ancestry, etc. For each individual, urine samples were collected at 16 time points. These samples were subsequently measured by ^{1}H-NMR metabolomics according to the protocols described in chapters 5 and 6. The medical ethical committee of the Radboud University Medical Center in Nijmegen approved the study (registration number 2012/350).

[5] A cross-validation procedure was used to estimate *Q*-values for the samples that were used to train an SHM model. Additionally, the sample corresponding to dot B in the figure was not used to train models. For construction of the SHM models it was assumed that samples taken at different time points for the same individual were independent. Small to modest autocorrelations derived under stationarity at various lags seems to support this claim. In order to estimate these autocorrelations, the unevenly sampled time series were first evenly resampled using a spectral approach based on the Lomb periodogram [14].

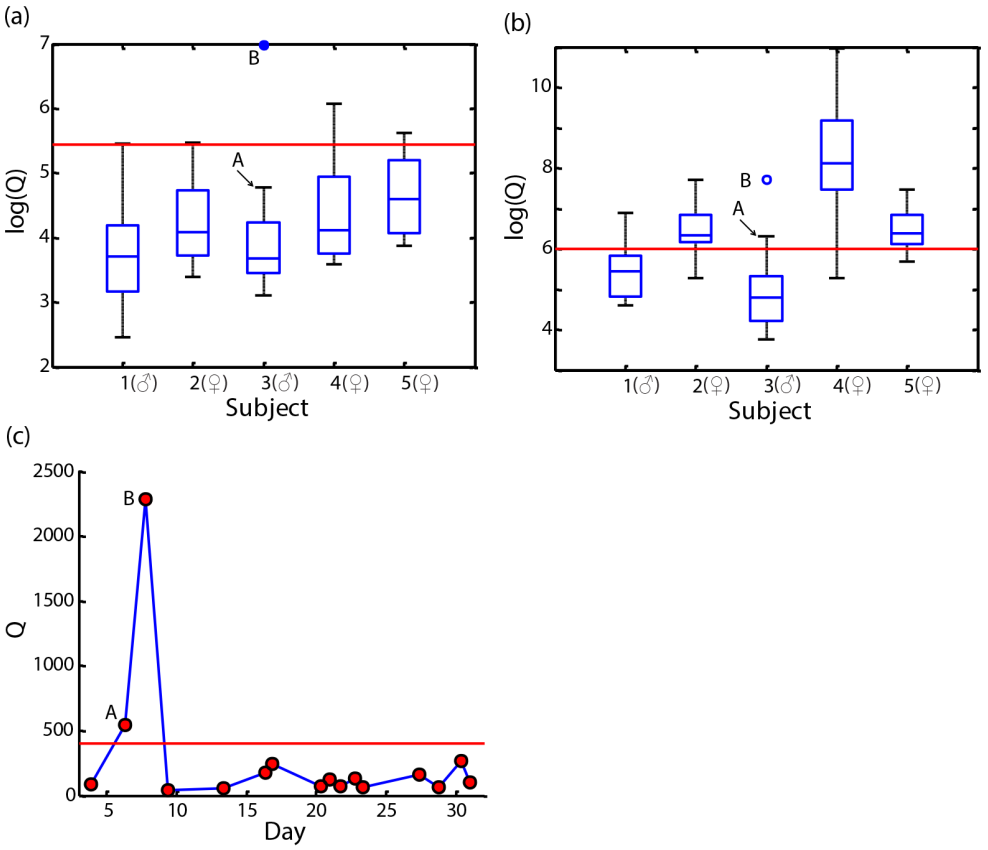

Figure 7.3 Analysis of the pilot project data with two different SHM models, namely **(a)** a population SHM model and **(b, c)** a personalized SHM model[5]. Similar to chapter 5, the *Q*-statistic was used to compare the metabolic profile of a sample to that of healthy controls in a multivariate manner where high *Q*-values indicate large differences. In panel **(a)** measurements of all individuals were used as healthy controls to train the SHM model and define NOC for the population (red horizontal line), where log(*Q*) indicates the natural logarithm of *Q*. The NOC matched the population well, but was, for example, too high for individual 3. In panels **(b, c)** measurements of individual 3 were used to construct the SHM model and define personalized NOC for that individual. The NOC were clearly more optimal for individual 3 compared to the population model. Additionally, the high *Q*-values of subjects 1, 2, 4, and 5 suggested that the personalized model described unique normal metabolic variation in individual 3; unlike the population model many normal measurements of the other individuals were clearly outside the NOC. These observations suggest that the use of personalized NOC would allow for earlier and better detection of subtle metabolic abnormalities in an individual. Subject 3 became ill during the test period (the flu) In panel **(c)** his *Q*-values are shown for all 16 time points with respect to the personalized NOC. It is interesting to see that an elevated *Q*-value (dot A) could already be observed 1.5 day before the individual felt sick or any clinical signs or symptoms of the flu became apparent (dot B). This was not observed when the SHM model was based on the population (panel **a**).

Many challenges remain for application of SHM for longitudinal health monitoring. For example, the optimal sampling frequency (days, weeks, months) needs to be estimated. The influence of natural shifts in metabolism, e.g. due to aging, on the performance of the SHM model also needs to be evaluated, and, if required, corrections for this need to be incorporated in the model. Additionally, it is important that the SHM method flags abnormal temporal events related to disease, but ignores abnormalities due to other environmental factors such as diet or medication. These "unimportant" abnormalities need to be identified and a way to cope with them must be found, e.g. by ignoring the variables associated to them.

Due to the close connection to industrial process monitoring it is expected that many ideas for longitudinal analysis developed in this field can also be useful for SHM. The methods developed in chapters 5 and 6 are essentially multivariate extensions of the so-called Shewhart control chart, which is the most popular chart in industry [17]. This chart has sometimes difficulty to detect small process changes. Alternative charts to detect such changes have been developed. Examples include the exponentially weighted moving average (EWMA) and the cumulative sum (CUSUM) charts [17]. In contrast to the Shewhart chart, these methods do not only use the current observation at each time point, but also consider historical information. It has been shown that the EWMA and CUSUM charts may detect small changes more quickly compared to the Shewhart chart, but large changes more slowly. Since many metabolic disruptions are expected to be rather subtle it is interesting to explore the use of EWMA, CUSUM, and other techniques for SHM. Multivariate generalizations of these methods with regularization by PCA (similar to the approach in chapter 5) have been proposed in literature [17]. The model proposed in chapter 6 can also be extended for EWMA and CUSUM analysis. The Shewhart, EWMA and CUSUM approaches all assume that the observations in time are exchangeable (i.e. uncorrelated). However, observations in time series are usually correlated to each other. Ignoring such autocorrelation can have an enormous impact on the performance of the charts. Therefore, this issue should be investigated with SHM. A possible solution is to fit a time series model to the data or to use a dynamic PCA approach. [17]. Finally, we would like to remark that the proposed SHM models (as well as the rMANOVA model developed in chapter 4) rely on the normal distribution. However, it is unlikely that metabolomics data follows this distribution, e.g. because a metabolite concentration cannot be detected below the lower limit of detection of the analytical technique that is used to measure it. Control charts have been developed for non-normal data. For example, Chang et al suggested a simple scaling of the data by "weighted standard deviations" to correct for skewed populations [18]. This method is directly applicable to the SHM models developed in chapters 5 and 6. Due to the close connection between the SHM and rMANOVA models via the Mahalanobis distance, it is

expected that this scaling method can also be used to extend rMANOVA to the non-normal case.

7.2.4 Concluding remarks

This thesis has focused on development of data analysis methods for high-dimensional data such as metabolomics data. Although work mainly centered on ^{1}H-NMR data, the proposed methods are general and also applicable to other data types and possibly much larger data sets. For example, rMANOVA and SHM hold great promise in "next generation" untargeted metabolomics where around 10000 signals derive from a single blood sample by techniques such as LC-MS using a quadrupole-time-of-flight mass spectrometer. Clearly, data analysis is indispensable for interpretation of such large data. Similarly, the techniques may also be used for analysis of data from (a combination of) other relevant omics technologies such as genomics, transcriptomics, glycomics, or proteomics.

Due to the increasing dimensionality of data sets, further development of the regularization approaches that were reviewed and developed in this thesis is crucial. The combination of rMANOVA with variable selection via e.g. a penalized approach as used in chapter 6 seems an interesting avenue for further research. Similarly, the sparse SHM model developed in chapter 6 may benefit from the regularized estimator of the covariance matrix that was used in chapter 4.

Independent of what kind of regularization is applied it is important to keep in mind that most untargeted metabolomics studies are explorative in nature due to the low number of samples that are considered. Obviously, subsequent validation of e.g. a discovered biomarker on larger populations is required to truly to assess its relevance and reliability before it can be used as e.g. a diagnostic tool [19]. This is also true for the proposed SHM approach.

References

1. Héberger, K., *Sum of ranking differences compares methods or models fairly.* TrAC Trends in Analytical Chemistry, 2010. **29**(1): p. 101-109.
2. Héberger, K. and K. Kollár-Hunek, *Sum of ranking differences for method discrimination and its validation: comparison of ranks with random numbers.* Journal of Chemometrics, 2011. **25**(4): p. 151-158.
3. Kollár-Hunek, K. and K. Héberger, *Method and model comparison by sum of ranking differences in cases of repeated observations (ties).* Chemometrics and Intelligent Laboratory Systems, 2013. **127**: p. 139-146.
4. Kalivas, J.H., K. Héberger, and E. Andries, *Sum of ranking differences (SRD) to ensemble multivariate calibration model merits for tuning parameter selection and comparing calibration methods.* Analytica chimica acta, 2015. **869**: p. 21-33.
5. Berk, M., T. Ebbels, and G. Montana, *A statistical framework for biomarker discovery in metabolomic time course data.* Bioinformatics, 2011. **27**(14): p. 1979-1985.
6. Smilde, A., et al., *Dynamic metabolomic data analysis: a tutorial review.* Metabolomics, 2010. **6**(1): p. 3-17.
7. Searle, S.R., *Linear models*. Wiley series in probability and mathematical statistics. 1971, New York, United States of America: John Wiley & Sons, Inc.
8. Mardia, K.V., J.T. Kent, and J.M. Bibby, *Multivariate analysis*. 1979: Academic Press.
9. Schelldorfer, J., L. Meier, and P. Bühlmann, *Glmmlasso: an algorithm for high-dimensional generalized linear mixed models using ℓ1-penalization.* Journal of Computational and Graphical Statistics, 2014. **23**(2): p. 460-477.
10. An, X., Q. Yang, and P.M. Bentler, *A latent factor linear mixed model for high-dimensional longitudinal data analysis.* Statistics in medicine, 2013. **32**(24): p. 4229-4239.
11. Kankainen, M., et al., *MPEA—metabolite pathway enrichment analysis.* Bioinformatics, 2011. **27**(13): p. 1878-1879.
12. Kamburov, A., et al., *Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA.* Bioinformatics, 2011. **27**(20): p. 2917-2918.
13. Wishart, D.S., *Advances in metabolite identification.* Bioanalysis, 2011. **3**(15): p. 1769-1782.
14. Chen, R., et al., *Personal Omics Profiling Reveals Dynamic Molecular and Medical Phenotypes.* Cell, 2012. **148**(6): p. 1293-1307.
15. Chen, R. and M. Snyder, *Promise of personalized omics to precision medicine.* Wiley Interdisciplinary Reviews: Systems Biology and Medicine, 2013. **5**(1): p. 73-82.
16. Li-Pook-Than, J. and M. Snyder, *iPOP goes the world: integrated Personalized Omics Profiling and the road towards improved health care.* Chemistry & biology, 2013. **20**(5): p. 660-666.
17. Bersimis, S., S. Psarakis, and J. Panaretos, *Multivariate statistical process control charts: an overview.* Quality and Reliability Engineering International, 2007. **23**(5): p. 517-543.
18. Chang, Y.S. and D.S. Bai, *A multivariate T2 control chart for skewed populations using weighted standard deviations.* Quality and Reliability Engineering International, 2004. **20**(1): p. 31-46.
19. Poste, G., *Bring on the biomarkers.* Nature, 2011. **469**(7329): p. 156-157.

Samenvatting

In de huidige gezondheidszorg zijn diagnostiek en behandeling meestal gericht op de "modale" patiënt. Dit heeft tot gevolg dat diagnostiek en behandeling succesvoller kunnen zijn voor de ene patiënt dan voor de andere patiënt. Met de snelle ontwikkelingen op het gebied van hoog-dimensionale analytische technologieën hoopt men hier verbetering in te brengen. Met deze technieken kan uiterst gedetailleerde informatie van een patiënt worden verkregen op moleculair niveau. Verwacht wordt dat dit zal resulteren in aanzienlijke vooruitgang in de gezondheidszorg. Allereerst kan een beter begrip van biologische processen worden verkregen, waarvan de mechanismen nu nog onduidelijk zijn. Verder kan met de nieuwe gedetailleerde kennis een meer persoonlijke gezondheidszorg worden gerealiseerd, waarbij diagnostiek en behandeling worden afgestemd op unieke moleculaire karakteristieken van een individuele patiënt.

Zoals beschreven in **hoofdstuk 1** was het doel van het onderzoek beschreven in dit proefschrift om verschillende problemen in de analyse van hoog-dimensionale data voor gepersonaliseerde gezondheidszorg op te lossen of nader tot een oplossing te brengen. Hierbij heeft het onderzoek zich vooral gericht op de analyse van zogenaamde "untargeted metabolomics" experimenten waar hoog-dimensionale technieken worden gebruikt voor het meten van de concentratie van een groot aantal kleine moleculen (metabolieten) in een monster, bijvoorbeeld urine of bloed. Verwacht wordt dat deze technieken steeds belangrijker worden voor gepersonaliseerde gezondheidszorg omdat concentraties van metabolieten in een monster zeer gedetailleerde informatie over het fenotype van een patiënt verschaffen. Dergelijke experimenten en technieken resulteren echter in zeer ingewikkelde data. Op dit moment schieten de mogelijkheden om deze data te analyseren te kort. Ontwikkeling van nieuwe analyse methoden is noodzakelijk.

Voorbewerking van data is een cruciale eerste stap in data analyse om ongewenste variatie in de data, zoals verschuivingen van de basislijn of piekpositie, te verwijderen. Dergelijke artefacten kunnen de relevante biologische informatie in de data volledig maskeren. Op basis van analyse van een "Fourier transform – Infrared benchmark data set" wordt in **hoofdstuk 2** aangetoond hoe lastig het kan zijn om geschikte methoden voor data voorbewerking te kiezen voor een specifieke data set. Slechts enkele van de potentieel geschikte voorbewerkingsmethoden resulteerden in betere resultaten na analyse, terwijl in veel gevallen de methoden juist averechts werkten. Bovendien werd aangetoond dat de drie veelgebruikte strategieën voor de keuze van geschikte voorbewerkingsmethoden ernstige nadelen vertonen: ze vergen onoverkomelijk veel rekentijd of geven misleidende resultaten. Het is duidelijk dat efficiëntere strategieën moeten worden ontwikkeld voor de keuze van voorbewerking van hoog-dimensionale data.

De volgende stap is het opstellen van een statistisch model, bijvoorbeeld om groepen monsters met elkaar te vergelijken of om een ziekte te diagnosticeren. Zoals beschreven in **hoofdstuk 3** zijn veel traditionele statistische methoden niet direct toepasbaar op metabolomics data, of zijn hun resultaten onbetrouwbaar. Een manier om deze problemen te omzeilen is om in het model structuur aan de data op te leggen met als doel de variantie van parameterschatters te verlagen ten koste van een (hopelijk geringe) (toename in) onzuiverheid. Dit staat bekend als regularisatie. **Hoofdstuk 3** beschrijft recente regularisatie methoden die van nut kunnen zijn voor de analyse van metabolomics data. Meer specifiek wordt aandacht besteed aan geregulariseerde schatters van de covariantie matrix, aangezien deze matrix in veel statistische methoden een belangrijke rol speelt. Vervolgens wordt het gebruik van een "l1-norm constraint" beschreven voor het aanpassen van een model aan data in combinatie met selectie van variabelen. De regularisatie methoden beschreven in hoofdstuk 3 zijn breed toepasbaar en vormen de basis voor verdere statistische analyse zoals beschreven in **hoofdstukken 4, 5 en 6**.

De eerste toepassing van regularisatie wordt beschreven in de context van analyse van metabolomics data die verkregen is volgens een experimentele proefopzet. Dergelijke experimenten zijn belangrijk voor gepersonaliseerde gezondheidszorg om de invloed van specifieke factoren en interacties op het metabolisme te onderzoeken om een beter begrip van de onderliggende biologische processen te verkrijgen. In **hoofdstuk 4** wordt bijvoorbeeld het effect van verschillende diëten op het metabolisme van metabool syndroom patiënten onderzocht. Multivariate variantie analyse (MANOVA) is een traditionele analyse methode voor data verkregen binnen een experimentele proefopzet. Deze methode kan echter doorgaans niet worden toegepast op data van metabolomics experimenten omdat een groot aantal variabelen (metabolieten of pieken) wordt gemeten op een betrekkelijk klein aantal monsters. Daarom is onlangs ANOVA simultaneous component analysis (ASCA) geïntroduceerd. In **hoofdstuk 4** wordt aangetoond dat binnen het ASCA model impliciet wordt verondersteld dat metaboliet concentraties ongecorreleerd (onafhankelijk onder normaliteit) zijn en dezelfde variantie hebben. Deze onrealistische veronderstellingen verminderen het onderscheidend vermogen (power) van ASCA en belemmeren de praktische interpretatie. Daarom wordt in **hoofdstuk 4** een verbeterde aanpak voorgesteld; geregulariseerde MANOVA (rMANOVA), MANOVA in combinatie met een recent geïntroduceerde geregulariseerde schatter van de (binnen-groep) covariantie matrix. In feite is rMANOVA een gewogen gemiddelde van ASCA en MANOVA. Met behulp van gesimuleerde data wordt aangetoond dat rMANOVA een hoger onderscheidend vermogen heeft dan ASCA onder een breed scala van mogelijke verschillen tussen groepen. Bij toepassing van rMANOVA en ASCA op urine ^{1}H-NMR metabolomics data, afkomstig uit

de eerder genoemde voedingsinterventie studie voor metabool syndroom patiënten, vindt rMANOVA meer verschillen tussen groepen dan ASCA.

In traditionele metabolomics studies worden groepen van monsters met elkaar vergeleken, zoals patiënten met een specifieke ziekte versus gezonde controle patiënten. Zoals betoogd in hoofdstukken 1, 5 en 6 kan deze aanpak onpraktisch zijn voor gepersonaliseerde gezondheidszorg en is een paradigmaverschuiving van populatie modellen naar meer gepersonaliseerde modellen vereist. In **hoofdstuk 5** wordt aangetoond dat ideeën uit de industriële proces controle nuttig zijn in dit opzicht. Op basis van deze ideeën wordt het concept van "statistical health monitoring" (SHM) geïntroduceerd. SHM kan worden gebruikt om automatisch de respons te detecteren (of misschien beter; de afwijking in de respons) die veroorzaakt wordt door een ziekte in een individuele patiënt. Dit wordt bereikt door de meetgegevens van een individuele patiënt (n = 1) te vergelijken met die van een groep controle monsters. Vervolgens identificeert het SHM model belangrijke verschillen tussen de meetgegevens van de controle groep en die van de patiënt. Een belangrijk voordeel is dat SHM alleen gegevens van controle monsters nodig heeft. Daarom is het SHM model geschikt om afwijkingen als gevolg van verschillende mogelijke ziektes (waaronder zeldzame ziektes) te diagnosticeren. In **hoofdstuk 5** wordt aangetoond dat verschillende "inborn errors of metabolism" (IEM, stofwisselingsziekten) kunnen worden gediagnosticeerd met ^{1}H-NMR metabolomics in combinatie met SHM. Diagnostiek van deze ziektes met traditionele populatie-modellen is lastig omdat rekening moet worden gehouden met een groot aantal ziekten en er meestal niet genoeg patiënt monsters beschikbaar zijn.

SHM maakt gebruikt van de "Mahalanobis distance" (MD) in combinatie met regularisatie via principale componenten analyse (PCA) om de gegevens van een patiënt met die van controles te vergelijken. SHM is ook toepasbaar wanneer het aantal controle samples kleiner is dan het aantal variabelen in de data set, vanwege de PCA regularisatie. Een nadeel is echter dat deze aanpak kan zorgen voor wat in de industriële proces controle als het "smearing effect" wordt omschreven. In feite een gevolg van het feit dat de associatie tussen variabelen voor zieke en gezonde patiënten niet noodzakelijk dezelfde is. Daarom wordt in **hoofdstuk 6** een alternatieve SHM aanpak geïntroduceerd. Bij deze aanpak wordt de MD geregulariseerd via variabele selectie in plaats van via dimensie reductie. Variabele selectie wordt gerealiseerd door de MD te schatten onder een "l1-norm constraint". Op basis van gesimuleerde data wordt aangetoond dat het resulterende sparse SHM model beter verschillen tussen controles en een patiënt kan aantonen. Het model wordt ook met succes toegepast op ^{1}H-NMR metabolomics data om verschillende IEM te diagnosticeren.

Data van hoog-dimensionale technologieën werpen veel problemen op voor een succesvolle data analyse. De belangrijkste bijdrage van het onderzoek beschreven in dit proefschrift bestaat uit een zorgvuldige evaluatie van huidige strategieën om de data voor te bewerken en de introductie van drie nieuwe statistische methoden, namelijk rMANOVA, SHM en sparse SHM, voor de analyse van metabolomics data voor gepersonaliseerde gezondheidszorg. De nieuwe methoden zijn met succes toegepast op urine ^{1}H-NMR metabolomics data om het effect van vier diëten op het metabolisme van metabool syndroom patiënten te toetsen, en om de gezondheidstoestand van individuele patiënten te inspecteren en verschillende zeldzame ziektes te diagnosticeren. De resultaten geven aan dat de nieuwe methoden grote potentie hebben voor gepersonaliseerde gezondheidszorg. Vervolg studies zijn echter vereist om resultaten verder te valideren en om de bruikbaarheid van de methoden in andere toepassingen te bestuderen.

Acknowledgements

I have thoroughly enjoyed my PhD research. This wouldn't have been possible without the help and support of many wonderful people at work and outside of work. Each of you has contributed to this thesis in your own, significant, way. I am deeply grateful to all of you.

First and foremost I would like to thank my (co-) supervisors Prof. L.M.C. Buydens, Prof. R.A. Wevers, Dr L. Blanchet and Dr U.F.H. Engelke. Lutgarde, je had al snel door dat chemometrie bij mij paste en wat ben ik blij dat ik ben in gegaan op je aanbod om bij jou te promoveren! Dank voor je enthousiaste, goede begeleiding en je betrokkenheid op zowel persoonlijk als professioneel vlak. Ik waardeer de ruimte die je me hebt gegeven om díe richting op te zoeken die ik zelf interessant vond binnen het onderzoek.
Ron, ondanks je drukke schema was je altijd benaderbaar en geïnteresseerd. Ik wil je bedanken voor de fijne begeleiding, het ontwikkelen van mijn kennis over metabolomics en je kritische en positieve blik op mijn werk en teksten.
Lionel, thank you for being my daily supervisor and a good friend. I really appreciate that I could always drop in your office to brainstorm or to discuss some "crazy" ideas or (more importantly?) other things in life than algebra. Thank you for finding the right words to motivate me when needed.
Udo, je hebt veel tijd gestoken in het meten van alle data voor hoofdstukken 4 – 7 van dit proefschrift. Mijn dank daarvoor en ook voor je bereidwilligheid om mijn vragen over de NMR spectra te beantwoorden; ik kon altijd bij je terecht en waardeer je positiviteit.

Besides my supervisors, I would also like to thank Prof. A.J. van Gool, Prof. E.R. van den Heuvel, and Dr R. Wehrens, the rest of my thesis committee, for taking the time to thoroughly read this manuscript. I hope it was a good read.
Alain, ik wil je ook bedanken voor de tijd die je hebt vrijgemaakt om mijn onderzoek en het vervolg van mijn carrière te bespreken.
Ron, dank voor de aanstekelijke manier waarop je de vakken statistiek en chemometrie doceerde, waarmee je me enthousiast hebt gemaakt voor dit onderwerp.

My sincere thanks goes out to Prof. R. Bro for giving me the opportunity to visit his department for 3 months and discover the amazing city of Copenhagen. To all my fellow "temporary colleagues" Jone, Idoia, Silvia, Silvia, Sonia, Andreas, Jochem and Jose; thank you for the great time and for introducing me to the Mediterranean way of life!

Heel veel dank aan al mijn collega's van Analytische Chemie: Jeroen, Geert, Brigitte, Jan, Ewa, Priyanka, Brigitte, Marietta, Agnieszka, Mike, Gerjen, Tom, Thahn, Nikola, Yajuan en Leslie. Bedankt voor de fijne samenwerking en gezelligheid op de afdeling. Mede door jullie

is mijn promotie een tijd om met veel plezier op terug te kijken. Ik zal de spelletjesavonden met veel te veel (ongezond) eten niet snel vergeten!
Jeroen, ik wil je heel erg bedanken voor je enthousiasme en de hulp bij het vinden van mijn nieuwe baan in Engeland!
Brigitte, dank voor alle ondersteuning rondom mijn proefschrift, ook nu ik in Engeland werk.
Ewa and Marietta, you were the best office mates I could have wished for. I have really appreciated your company, our conversations about anything and your support. Dzięki; ευχαριστίες.

Ik wil Bart, Bayan, Ingrid en Sikko, "mijn" studenten, heel erg bedanken dat ik jullie mocht begeleiden tijdens jullie Bachelor en Master stages. Ik vond het erg leuk om met jullie samen te werken en jullie te zien groeien op wetenschappelijk gebied.

In het dagelijks leven buiten de RU wordt ik ook omringd door veel fijne vrienden.
Mireille, dank dat je mijn paranimf wilt zijn terwijl je zelf ook druk bent met je eigen promotie. Ik ben heel blij met onze vriendschap en kijk alweer uit naar het volgende etentje (sushi maar weer?).
Jurn ik voel me vereerd dat jij straks als paranimf naast mij zal staan op de verdediging. Dat is ook wel terecht na veel etentjes, bioscoopbezoeken, vakanties en twee keer de vierdaagse. Bedankt!
Aan al mijn andere "vierdaagse maatjes", Bram, Michelle, Eline, Michiel, en Roel dank voor de gezelligheid tijdens de vele kilometers. Ook daarbuiten kon ik de afgelopen jaren vaak bij jullie terecht voor lekker eten en een fijn gesprek, wat voor de broodnodige afleiding zorgde. Dat waardeer ik zeer.
Arwin, ook wij hebben in Kopenhagen heel wat kilometers afgelegd op zoek naar de afterparty van de "Kulturnatten". Inmiddels is je knie gelukkig weer helemaal in orde ☺. Ik vind het altijd weer leuk om met je af te spreken. Je enthousiasme en gedrevenheid zijn inspirerend.
Jan en Cynthia, ik kan bij jullie altijd terecht voor een gezellige avond. Het spelen van allerlei bordspellen met jullie is altijd leuk en competitief. Jan, dank voor alle gezelligheid tijdens het delen van een kamer op meerdere conferenties, het brainstormen wanneer ik vast zat met een chemometrisch probleem, en de fijne samenwerking tijdens al onze onderwijs activiteiten.
Simeon, ondanks dat je me tijdens het thermo prakticuum door een "klein foutje" bijna verzoop zijn we al bijna 10 jaar vrienden. Ik ben blij dat de vriendschap er sinds onze studie tijd niet minder op is geworden en ik hoop dat dit nog lang zo blijft. Veel geluk en succes in Naarden!

Rob, dankjewel voor alle gezellige bezoeken aan Doornroosje, Merleyn en Lowlands; en voor de introductie in het verzamelen van vinyl.
Rolf, ik had nooit verwacht zo'n goede vriend aan de andere kant van de wereld in Thailand te ontmoeten. Ik kijk uit naar onze volgende citytrip samen!
Priyanka, I have really appreciated your great company over the years. Thank you for the unforgettable holiday to Spain. I hope you have recovered from the loss of your Samsung phone and the fact that you now have to do with just a mere iPhone ☺.
Lionel and Agi, thank you for inviting me to Poland and for being good friends over the years. I am inspired by the way how you manage to combine family with science.
Seth, hoewel we elkaar niet meer zo vaak spreken als op de middelbare school beleef ik altijd veel plezier aan je verhalen over wielrennen en fotografie. Ik ben blij dat ik onze fietstocht in Toscane heb overleefd!

Tot slot mijn familie.
Oma, bedankt voor alle keren dat je zo lekker voor me hebt gekookt, je interesse, en de gezellige gesprekken door de jaren heen. Alle mooie verhalen over opa (Frits) zijn een inspiratie in mijn werk.
Opa, ook bij jou en oma (Nel) kan ik al mijn hele leven terecht om even helemaal te ontspannen. Bedankt voor alle super vakanties toen Josien en ik nog jong waren. Ook nu we tegenwoordig niet meer samen op de fiets kunnen stappen kom ik nog altijd graag op bezoek!
Heiko en Doki, bedankt voor de ontspanning die ik op Terschelling heb kunnen genieten in de laatste periode van de promotie. Ik vind het heel fijn dat ik me zo thuis kan voelen bij mijn "nieuwe familie".
Josien, ik had me geen betere band met mijn zus kunnen wensen. Ik ben je dankbaar voor het luisterend oor als ik daar behoefte aan had. Wat bijzonder dat we nu op hetzelfde moment onze promotie afronden! Matthijs, bedankt voor de gezellige avonden. Ik vind het altijd leuk om met je te brainstormen over je creatieve ideeën. Sophie, hoewel je nog geen bijdrage hebt kunnen leveren aan het proefschrift zelf ben ik je dankbaar voor de levendigheid die je in onze familie hebt gebracht.
Mama en papa door jullie opvoeding ben ik geworden wie ik nu ben, maar wie had ooit kunnen denken dat we uiteindelijk zelfs in hetzelfde vakgebied werkzaam zouden zijn! Van kinds af aan hebben jullie me altijd gestimuleerd om nieuwsgierig te zijn en te doen wat ik leuk vind. Bedankt voor jullie liefde en onvoorwaardelijke steun. Met trots draag ik dit proefschrift aan jullie op.
Een absoluut hoogtepunt van een werkdag is als ik jou, Martine, met een grote glimlach op je gezicht voor het raam zie staan wanneer ik thuis kom. Lieve schat, bedankt voor alle steun bij het afmaken van dit proefschrift. Ik ben heel gelukkig dat je elke dag bij mij wilt zijn.

LIST OF PUBLICATIONS

Published articles:

Engel, J., Blanchet, L., Buydens, L. M. C., & Downey, G. (2012). Confirmation of brand identity of a Trappist beer by mid-infrared spectroscopy coupled with multivariate data analysis. *Talanta*, *99*, 426-432.

Smeenk, N. J., **Engel, J.**, Mulder, P., Bauhuis, G. J., Bissels, G. M. M. W., Schermer, J. J., ... & Kelly, J. J. (2013). Arsenic formation on GaAs during etching in HF solutions: relevance for the epitaxial lift-off process. *ECS Journal of Solid State Science and Technology*, *2*(3), 58-65.

Engel, J.[#], Gerretzen, J.[#], Szymańska, E., Jansen, J. J., Downey, G., Blanchet, L., & Buydens, L. M. C. (2013). Breaking with trends in pre-processing?. *TrAC Trends in Analytical Chemistry*, *50*, 96-106.

Engel, J., Blanchet, L., Engelke, U. F., Wevers, R. A., & Buydens, L. M. C. (2014). Towards the Disease Biomarker in an Individual Patient Using Statistical Health Monitoring. *PloS one*, *9*(4).

Szymańska, E., Gerretzen, J., **Engel, J.**, Geurts, B., Blanchet, L., and Buydens, L. M. C. (2015). Chemometrics and qualitative analysis have a vibrant relationship. *TrAC Trends in Analytical Chemistry, 69*, 34-51.

Engel, J.[#], Postma, G. J.[#], van Peufflik, I., Blanchet, L., & Buydens, L. M. C. (2015). Pseudo-sample trajectories for variable interaction detection in Dissimilarity Partial Least Squares. *Chemometrics and Intelligent Laboratory Systems*, *146*, 89-101.

Engel, J., Blanchet, L., Bloemen, B., van den Heuvel, L. P., Engelke, U. H. F., Wevers, R. A., & Buydens, L. M. C. (2015). Regularized MANOVA (rMANOVA) in untargeted metabolomics. *Analytica chimica acta*, *899*, 1-12 (Featured article).

Reinen, J., Postma, G., Tump, C., Bloemberg, T., **Engel, J.**, Vermeulen, N. P., ... & Honing, M. (2016). Application of a cocktail approach to screen cytochrome P450 BM3 libraries for metabolic activity and diversity. *Analytical and Bioanalytical Chemistry*, *408 (5),* 1425-1443.

[#] These authors contributed equally to this study.

Submitted articles:

Engel, J., Blanchet, L., Engelke, U. F. H., Wevers, R. A., & Buydens, L. M. C. A Novel Approach to Diagnosis and Follow-up of Individual Patients by Sparse Modeling.

Geurts, B. P., **Engel, J.**, Rafii, B., Blanchet, L., Jansen, J. J., & Buydens, L. M. C. Improving high-dimensional data fusion by exploiting the multivariate advantage.

Singh, P., **Engel, J.**, Jansen, J. J., de Haan, J., & Buydens, L. M. C., Dissimilarity based partial least squares (DPLS) for genomic prediction from whole genome SNP data.

Di Guida, R., **Engel, J.**, Allwood, J. W., Weber, R. J. M., Jones, M. R., Sommer, U., Viant, M. R., & Dunn, W. B. Non-targeted UHPLS-MS metabolomics data processing workflow: A comparative investigation of normalization, missing value imputation, transformation and scaling.

Southam, A. D.[#], Weber, R. J. M.[#], **Engel, J.**, Jones, M. R., & Viant, M. R. High-Resolution Nanoelectrospray Direct Infusion Mass Spectrometry-based Metabolomics and Lipidomics: A Complete Workflow.

Davis, J. E., Viant, M. R., Duffy, C., **Engel, J.**, Sebire, M., & Katsiadaki, I. Application of Chemcatcher passive sampler to characterise the fish exometabolome.

Euceda, L. R., Haukaas T. H., Giskeødegård, G. F., Vettukattil, R., **Engel. J.**, Silwal-Pandit, L., ... & Bathen, T. F. Evaluation of metabolomics changes during neoadjuvant chemotherapy combined with bevacizumab in breast cancer using MR spectroscopy.

Curriculum vitae

Jasper Engel was born in 1984 in Gouda, The Netherlands. He graduated from secondary school (VWO) in 2004. Following this, he obtained his Bachelor's degree in Chemistry (*cum laude*) at Radboud University Nijmegen, the Netherlands in 2008. In 2011 Jasper received his Master's degree in Chemistry (*cum laude*) from the same university. During his graduation project at Applied Materials Science, Radboud University Nijmegen, he investigated etching of GaAs wafers in hydrofluoric acid by microscopic techniques, profilometry and X-ray photoelectron spectroscopy. The results of this study were published in the *ECS Journal of Solid State Science and Technology*. A second research project was carried out at Teagasc Ashtown Food Research Centre, Ireland, where he used chemometric techniques in concert with infrared spectroscopy in a food authentication study. The results of this study were published in *Talanta*. In 2011 Jasper was awarded the Unilever Research Prize for the quality of the research performed in both projects. During his Master's, Jasper also performed several extracurricular activities including teaching and organizing the scientific and cultural program of a study tour to Argentina.

In January 2011 Jasper started his PhD research at Radboud University Nijmegen under supervision of Prof. L.M.C. Buydens and Prof. R.A. Wevers. His research focused on the development of novel data analysis approaches for metabolomics and personalized health care. In the second year of his PhD Jasper visited the Department of Food Science, University of Copenhagen, Denmark for three months where he studied the application of multiway methods for analysis of COSY NMR metabolomics data. During his PhD, Jasper presented his work at several conferences and was twice awarded a prize for best poster presentation. Jasper was invited to the graduate school pilot project of the Institute of Molecules and Materials. He was also involved in several courses on statistics and chemometrics at Bachelor's level and supervision of Bachelor's and Master's students during their research projects.

Since January 2015 Jasper has been appointed as Experimental Officer in biostatistics at NERC Biomolecular Analysis Facility – Metabolomics node, School of Biosciences, University of Birmingham, United Kingdom. Here, his research focuses on the development and application of data pre-processing and statistical modeling approaches for analysis of (environmental) metabolomics data. Additionally, Jasper lectures on data analysis for metabolomics in several courses provided by the Birmingham Metabolomics Training Facility. In the next years Jasper aims to continue to develop himself in chemical data science (chemometrics) focusing on the development of novel approaches with applications in metabolomics, (personalized) medicine, and the environmental sciences.